



Report NAVTRAEQUIPCEN 74-C-0063-1

FC.
12

DEVELOPMENT AND EVALUATION OF TRAINEE PERFORMANCE MEASURES
IN AN AUTOMATED INSTRUMENT FLIGHT MANEUVERS TRAINER

Canyon Research Group, Inc.
32107 Lindero Canyon Road, Suite 123
Westlake Village, California 91361

MAY 1976

Final Report for Period 17 January 1974 - 17 October 1975

DOD DISTRIBUTION STATEMENT

Approved for public release;
distribution unlimited

Prepared for

Human Factors Laboratory
NAVAL TRAINING EQUIPMENT CENTER
Orlando, Florida 32813

DDC
RECEIVED
MAY 18 1976
B

NAVAL TRAINING EQUIPMENT CENTER
ORLANDO, FLORIDA

5/1 391185

NAVTRAEQUIPCEN 74-C-0063-1

GOVERNMENT RIGHTS IN DATA STATEMENT

Reproduction of this publication in whole or in part is permitted for any purpose of the United States Government.

ADDITIONAL FOR

BY: ☒ W/In Section ☒ BY

DOC ☐ Diff Section ☐

UNANNOUNCED

JUSTIFICATION

BY: ☒ AVAIL. MILITARY CODES

BY: ☒ AVAIL. AND/OR SPECIAL

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NAVTRAEQUIPO 74-C-0063-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE AND SUBTITLE Development and Evaluation of Trainee Performance Measures in an Automated Instrument Flight Maneuvers Trainer.	5. PERFORMING ORG. REPORT NUMBER	6. DATE OF REPORT & PERIOD COVERED Final Report - 17 JAN 1974 - 17 OCT 1975
7. AUTHOR(s) Donald Vreuls, A. Lee Wooldridge, Richard W. Obermayer, Robert M. Johnson, Don A. Norman and Ira Goldstein	8. CONTRACT OR GRANT NUMBER(s) N61339-74-C-0063 NEW	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Canyon Research Group, Inc. 32107 Lindero Canyon Road, Suite 123 Westlake Village, California 91361	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 3754-02-P01	
11. CONTROLLING OFFICE NAME AND ADDRESS Human Factors Laboratory Naval Training Equipment Center Orlando, Florida 32813	12. REPORT DATE 17 Oct 1975	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	13. NUMBER OF PAGES 112	
	15. SECURITY CLASS. (of this report) UNCLASSIFIED	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES Phase I of this research (17 January 1974 - 17 July 1974) was sponsored by Advance Projects Research Agency, ARPA Order 2310, Amendment 1, Program Code 4W10.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Performance Measurement Measure Selection Analysis Training Performance Multiple Discriminant Analysis Automated Flight Simulator Computerized Measurement System Training Measurement System Evaluation Selection of Measures		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A simulator study was conducted to improve training performance measurement selection methods, apply the results to an automated flight training system and conduct an evaluation of resulting measurement during automated training of four instrument flight maneuvers. Empirical methods were used to select from an analytically derived set, those measures which had the ability to discriminate		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

391185 ✓

✓B

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

between early and later training performance. The multiple discriminant model emerged as the best technique, but the algorithm for its use was highly modified. The automated trainer was then modified to operate on three measurement subsystems, (1) the original scoring algorithm, (2) the measures and weighting coefficients based on multiple discriminant analysis results, and (3) the original scoring algorithm using measured normative data.

Resulting measurement was evaluated by automatically trained three matched groups of five civilian pilots each with the result that time-to-train was reduced 34-40% for pilots training with empirically derived measures over the original scoring algorithm. It was recommended that data collection at an operational site be undertaken to verify the methods and to produce information that might lead to a measurement specification for future devices. Recommendations concerning the design of adaptive logics were made.

REVISION for

DATE 11/11/80

UNCLASSIFIED

JUSTIFIED

BY [Signature]

CODES

ORIGINAL

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

SUMMARY

The development, implementation and empirical evaluation of a man-vehicle training performance measurement method is reported herein. Initial work by Vreuls, Obermayer, Lauber and Goldstein (1973) emphasized the development of a descriptive structure for obtaining measurement in a man-vehicle training situation, starting with analytical specification of measures. Noticing certain deficiencies in measurement produced by analytic methods alone, the next effort (Vreuls, Obermayer and Goldstein, 1974) centered on the initial exploration of empirical measure selection techniques to be used in conjunction with the descriptive model.

Phase I of the present effort concentrated on further refinement of measure selection methods and application of those methods to empirically derived data for the purpose of recommending measures for use in automated instrument flight simulator training. The criterion for selection of analytically defined measurement was that each measure had to be able to discriminate between early and later training. Tests of significant changes for singular measures, correlations between measures, multiple discriminant analyses and canonical correlation analyses were explored. A modified form of the multiple discriminant analysis appeared most suitable for the purpose. Measures, weighting coefficients, and measurement start and stop conditions resulted from analysis of data obtained from the training of 12 pilots on four instrument maneuvers.

Phase II focused on the insertion of Phase I measurement results into the automated Instrument Flight Maneuvers flight simulator (TRADEC/IFM) at NAVTRAEQUIPCEN in real time, and the development of a rationale to map the new and somewhat different measure sets into an adaptive logic, or task scheduler which was designed to accept slightly different information. IFM was modified to operate with three measurement subsystems, (1) the original scoring algorithm, (2) the recommended measures from Phase I, and (3) the original scoring algorithm modified on the basis of normative data obtained in Phase I.

Resulting measurement subsystems were evaluated in Phase III by automatically training three matched groups of five civilian pilots each. The time-to-train each group to the same performance criteria was reduced 34-40% for both empirically derived measure groups (2 and 3 above) over the original, analytically defined measurement algorithm. The discriminant measures appeared to be sensitive to piloting technique and provide more reliable performance feedback. Also, the discriminant model appeared to have potential for growth to higher efficiency levels than reported because of its ability to select and properly weight important student variables along with system performance. Potentially serious inefficiencies with linear, single score adaptive logics were observed and discussed.

NAVTRAEQUIPCEN 74-C-0063-1

The results were encouraging enough to recommend that data collection at an operational training site be undertaken to verify the measure selection methods within the context of a military flight training environment, and to produce data which might lead to eventual measurement specification for future training devices for the class of aircraft and maneuvers flown. Recommendations were made also for improvement of adaptive logics similar to IFM and for a relatively inexpensive study that might resolve adaptive logic inefficiency and provide valuable guidance to designers.

PREFACE

The authors wish to gratefully acknowledge the assistance and insight of four people who contributed to the success of this work. Mr. Hal McKinney and Mr. Jerry Diddle of NAVTRAEQUIPCEN kept the TRADEC operational and provided assistance during data collection above and beyond normal duty. Dr. Robert Breaux, NAVTRAEQUIPCEN Human Factors Laboratory provided many hours of valuable guidance on multivariate statistical issues; he pointed to the solution of the repeated measures problem in the discriminant analysis that had plagued the research community for some time. Dr. David G. Weinman, Department of Statistics, Hollins College, Virginia, also provided assistance on statistical issues while he was a summer employee of NAVTRAEQUIPCEN.

A separate report, more thoroughly addressing the statistical issues encountered in this study, is planned for publication in early 1976 under the following authorship, report number and approximate title:

Wooldridge, A.L. Breaux, R., and Weinman, D. "Statistical Issues in the Use of Multivariate Methods for Selection of Flight Simulator Performance Measures."
NAVTRAEQUIPCEN 75-C-0091-1.

The new military standard reporting format does not allow convenient front page identification of authors' organizations, where more than one organization is involved. Mr. Don A. Norman and Mr. Ira Goldstein performed technical contributions to the effort as NAVTRAEQUIPCEN employees. The services of Mr. Richard W. Obermayer were provided on subcontract with Manned Systems Sciences, Inc., 8949 Reseda Blvd., Suite 214, Northridge, California 91324. The services of Mr. Robert M. Johnson were provided on subcontract with Appli-Mation, Inc., 1000 Woodcock Road, Suite 174, Orlando, Florida 32813.

Phase I of this research was sponsored by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Naval Training Equipment Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Advanced Research Projects Agency of the United States.

TABLE OF CONTENTS

<u>Section</u>	<u>Page</u>
I. INTRODUCTION AND TECHNICAL SUMMARY	11
Measure Selection Summary	12
Measurement Implementation Summary	13
Measurement Evaluation Summary	15
II. MEASURE SELECTION METHOD	17
Measure Selection Process Summary	17
Apparatus	18
Participants	18
Task	19
Procedure	20
Experimental Design	20
Measurement	24
Measure Selection Analyses	31
III. MEASURE SELECTION RESULTS AND DISCUSSION	37
Means and t-Tests	37
Equivalent Measures	37
DISCRIM SELECT	40
Recommended Measures and Weights	41
Discussion	47
IV. MEASUREMENT IMPLEMENTATION	51
Apparatus	51
Training Course	51
Original IFM Performance Scoring	53
Original IFM Adaptive Logic	56
Mapping New Measures into Existing Adaptive Logic	56
Measurement Implementation	62
System Test Procedures	63/64
V. MEASUREMENT SYSTEM EVALUATION	65
Method	65
Results	66
Discussion	71
VI. CONCLUSIONS	75
Measure Selection	75
Measure Implementation	76
Measure Evaluation	77

NAVTRAEQUIPCEN 74-C-0063-1

TABLE OF CONTENTS - cont

<u>Section</u>	<u>Page</u>
VII. RECOMMENDATIONS	80
REFERENCES	83
APPENDIX A - RAW DATA AND MEASUREMENT FUNCTIONS AND TRANSFORMS AVAILABLE IN CURRENT MEASUREMENT PROGRAMS	85
APPENDIX B - CANDIDATE MEASURE MEANS AND t-TESTS BY MANEUVER	89
APPENDIX C - EQUIVALENT MEASURES BY MANEUVER	95
APPENDIX D - IFM PROGRAM MODIFICATIONS TO INCORPORATE PERFORMANCE MEASUREMENT TECHNIQUES	100
APPENDIX E - TYPICAL TRAINING PROFILES	107
APPENDIX F - COMMENTS ON AUTOMATED TRAINING SYSTEM DESIGN	111

NAVTRAEQUIPCEN 74-C-0063-1

LIST OF TABLES

<u>Table No.</u>		<u>Page</u>
1	EXPERIMENTAL DESIGN	21
2	CANDIDATE MEASURES FOR MANEUVER 1, STRAIGHT AND LEVEL	26
3	CANDIDATE MEASURES FOR MANEUVER 2, CLIMBS AND DESCENTS	27
4	CANDIDATE MEASURES FOR MANEUVER 3, LEVEL TURNS .	28
5	CANDIDATE MEASURES FOR MANEUVER 4, CLIMBING AND DESCENDING TURNS	29
6	CANDIDATE MEASURES FOR MANEUVER 4, CLIMB OR DIVE AND TURN REVERSAL	30
7	NUMBER OF MEASURES SELECTED BY t-TESTS	38
8	NUMBER OF EQUIVALENT MEASURES	38
9	NUMBER OF MEASURES REMAINING FOR MULTIVARIATE ANALYSES	39
10	NUMBER OF MEASURES IN EACH MINIMUM DISCRIMINATING SET	40
11	RECOMMENDED MEASURES AND WEIGHTS FOR MANEUVER 1, STRAIGHT AND LEVEL	42
12	RECOMMENDED MEASURES AND WEIGHTS FOR MANEUVER 2, CLIMBS AND DIVES	43
13	RECOMMENDED MEASURES AND WEIGHTS FOR MANEUVER 3, LEVEL TURNS	44
14	RECOMMENDED MEASURES AND WEIGHTS FOR MANEUVER 4, CLIMBING AND DIVING TURNS	45
15	MEANS AND STANDARD DEVIATIONS OF DISTRIBUTIONS IN DISCRIMINANT SPACE	46
16	MODIFIED SYLLABUS	52
17	ORIGINAL IFM PERFORMANCE BAND LIMITS	54
18	ORIGINAL IFM PARAMETERS SCORED	54
19	ORIGINAL IFM WEIGHTING COEFFICIENTS	55

NAVTRAEQUIPCEN 74-C-0063-1

LIST OF TABLES - cont

<u>Table No.</u>		<u>Page</u>
20	ORIGINAL IFM SCORING ALGORITHM	55
21	ORIGINAL IFM ADAPTIVE LOGIC	56
22	AVERAGE TOTAL WEIGHTED SCORES FOR USE IN NEW MEASUREMENT SYSTEM	58
23	UPPER BOUNDS OF NEGATIVELY WEIGHTED MEASURES . .	59
24	AVERAGE IFM SCORES FROM PHASE I	60
25	ADAPTIVE LOGIC FOR ALL SCORING SYSTEMS	61
26	TRAINEE DATA	67
27	RAW RESULTS	67
28	SIMPLE CORRELATIONS BETWEEN VARIABLES	68
29	MULTIPLE REGRESSION RESULTS	68
30	MEASUREMENT EVALUATION RESULTS	70
31	NUMBER OF RAW TRIALS TO COMPLETE EACH MANEUVER .	70
32	REAL TIME RAW DATA PARAMETERS FROM SIMULATOR . .	85
33	GLOSSARY OF START/STOP FUNCTIONS	86
34	GLOSSARY OF LOGICAL OPERATORS FOR COMBINING START/STOP FUNCTIONS	86
35	GLOSSARY OF TRANSFORMATIONS	87
36	AVERAGE MANEUVER 1 (STRAIGHT & LEVEL) MEASURES .	89
37	AVERAGE MANEUVER 2 (CLIMBS & DESCENTS) MEASURES .	90
38	AVERAGE MANEUVER 3 (LEVEL TURNS) MEASURES	91
39	AVERAGE MANEUVER 4, SEGMENT 2 (INITIAL CLIMB/DIVE TURN) MEASURES	92
40	AVERAGE MANEUVER 4, SEGMENT 3 (CLIMB/DIVE & TURN REVERSAL) MEASURES	93
41	AVERAGE MANEUVER 4, SEGMENT 4 (FINAL CLIMB/DIVE TURN) MEASURES	94

NAVTRAEQUIPCEN 74-C-0063-1

LIST OF TABLES - cont

<u>Table No.</u>		<u>Page</u>
42	MANEUVER 1 (STRAIGHT & LEVEL) EQUIVALENT	95
43	MANEUVER 2 (CLIMBS & DIVES) EQUIVALENT MEASURES .	96
44	MANEUVER 3 (LEVEL TURNS) EQUIVALENT MEASURES . .	97
45	MANEUVER 4, SEGMENT 2 (INITIAL CLIMB/DIVE TURN) EQUIVALENT MEASURES	98
46	MANEUVER 4, SEGMENT 4 (FINAL CLIMB/DIVE TURN) EQUIVALENT MEASURES	98
47	MANEUVER 4, SEGMENT 3 (CLIMB/DIVE TURN REVERSAL) EQUIVALENT MEASURES	99

LIST OF ILLUSTRATIONS

<u>Figure No.</u>		<u>Page</u>
1	DISCRIM SELECT FUNCTIONAL FLOW	34
2	TYPICAL GROUP 1 (OLD IFM) SUBJECT PERFORMANCE . .	108
3	TYPICAL GROUP II (DISCRIM) SUBJECT PERFORMANCE .	109
4	TYPICAL GROUP III (NORM IFM) SUBJECT PERFORMANCE	110

SECTION I

INTRODUCTION AND TECHNICAL SUMMARY

Measurement produces information which is needed for assessment of trainee performance, subsequent control of training and for training effectiveness evaluation. Improvements in training efficiency, and evaluation of training methods are quite dependent on improved measurement. Any device, system or process which is to control or evaluate training will be only as effective as its information sources.

In order to measure many of the complex dimensions of man-machine system training performance, the processing of large amounts of continuously varying information is required. Such measurement is beyond the capability of manual or simple measurement devices; it must be automated in order to produce information in time for effective control of training.

Automated measurement places severe demands on the definition of (a) fool-proof algorithms for determining the conditions during which measurement is to occur, and (b) measure sets which produce only the information necessary for effective use by the information receiving system. Too much information can overload the user; not enough information might reduce user effectiveness.

Historically, performance measures have been specified by analyses of knowledges, tasks, mission requirements and performance standards drawn from experience or consensus of experts. Analytically derived measurement is likely to include (a) different measures of the same or closely related behavior, (b) measures which may prove to be unimportant and (c) measurement based on oversimplified or inaccurate criteria. Although measurement development must start with a good analysis, empirical techniques are required to overcome analytic difficulties and reduce measurement to a small, efficient set. The reduction of analytically defined measures into a set which can be shown to have the desired properties is called the measure selection process herein.

Previous work has established and tested (a) a descriptive structure, or model, for obtaining measurement in man-machine training and (b) measure selection methods based on multivariate statistical models which evaluate the total set of measures taken together, and produce valuable weighting coefficients. This work led to the present three phase study to (a) refine the measure selection methods, (b) apply the results to an automated flight training system and (c) conduct tests and evaluations of the resulting measurement. Since this report is quite lengthy, a technical summary of the work is presented in the following pages of this section.

MEASURE SELECTION SUMMARY

The purpose of Phase I was to improve measure selection methods while developing measures for an experimental automated Instrument Flight Maneuvers (IFM) training system located at NAVTRAEQUIPCEN. The automated system was modified to control a measurement study (rather than automatically train). Data were collected on magnetic tape while 12 low-time pilots underwent 18, one-hour training sessions on four instrument flight maneuvers.

The resulting data were used for measure selection analyses at the conclusion of training. Initially, an average of 16 performance measures were produced for each maneuver and measure segment. Correlational analyses of redundant information reduced the average number of measures from 16 to 12. A multiple discriminant analysis was used to find the measures and weighting coefficients that would best describe the change in performance from early to later training; an average of six measures were found to be important for each measure segment. With the addition of specified outer loop measures the recommended set which averaged 9 measures could be weighted and summed into a single score, the discriminant function.

Canonical correlation analyses were explored also to uncover predictive relationships between measure sets early and late in training. They produced an average of seven measures per maneuver. They also produced asymmetrical predictive and criterion sets that were difficult to interpret and relate to the multiple discriminant analysis results. Since the multiple discriminant analysis can be interpreted as a form of prediction, and the results were difficult to bring together, the canonical correlation was omitted from further development.

STATISTICAL PROBLEM. Due to experimental design restrictions, four problems of a statistical nature arose because of our desire to use the multiple discriminant analysis. The first problem was that the mathematics of multivariate methods demand that there be more independent observations in each treatment group than unique variables. Although in each day there were 144 observations and only an average of 16 variables, the observations could not be considered "independent" because only 12 subjects were observed (12 times each day).

The second problem came from the underlying assumption in the derivation of multiple discriminant analysis that the various treatment groups be independent. Since each subject was measured in all of the treatment groups, the experimental design also failed the requirement of independent groups.

A third problem arose because we planned to use the weights derived from Phase I data in a subsequent application with a new group of students. The reliability of weighting coefficients from application to application has been questioned.

A fourth and final problem arose because the weighting coefficients for maneuvers flown with turbulence could not be determined with accuracy because there were no turbulence runs early in training to be paired with later turbulence runs for the discriminant analysis.

STATISTICAL PROBLEM SOLUTION. A method was derived from the literature to remove the components of variance due to both repeated observations within a group, and repeated observations between groups (the first two problems above). As can be seen in Section II, the method was similar to those used in univariate statistics for repeated observations.

It was discovered also in the literature that a technique for improving the predictive reliability of weighting coefficients had been developed for the multiple regression analysis. The method, called "Ridge Regression," incrementally adds a small bias to the diagonal of the intercorrelation matrix prior to multiple regression. As the bias is added, the weighting coefficients can be seen to asymptote to stable values. The third problem, weighting coefficient reliability, was solved by adding a small bias to the "within groups" matrix in the discriminant analysis (similar to a ridge regression). The results markedly changed the extreme values of certain coefficients without altering the power of the discrimination.

The fourth problem was alleviated by removing turbulence from the syllabus for subsequent implementation and evaluation phases.

A small mathematical controversy still lingers over the solution to the statistical problems of this study. These arguments are being researched and described in a separate study entitled, "Statistical Issues." These statistical issues were considered more or less fine tuning in their relation to the overall measurement system and not to have a large impact on the concept of the discriminant measurement system.

MEASUREMENT IMPLEMENTATION SUMMARY

The purpose of Phase II was the implementation of measures, weighting coefficients and conditional expressions to start and stop measurement (from Phase I) in the IFM system so that it could train in the automated mode with three measurement subsystems in the Phase III evaluation. The three subsystems to be used were (1) original IFM scoring, (2) scoring based on discriminant analysis results and (3) scoring based on the original IFM measures, corrected for measured performance norms.

One major technical challenge was to make the new measurement system operate in real time. The basic flight program required solutions of the aerodynamic equations every 50-milliseconds, and it took about 35-milliseconds to process the equations themselves. That left about 15-milliseconds to per-

form all of the existing IFM functions (which make the flight system an automated trainer) and to process the new measurement and measure start and stop functions, which were more complex (than the original). Modular and somewhat hierarchical software design, elimination of "nice-to-have" but unnecessary real-time performance plots, and rearrangement of background and foreground processing functions provided solution to the processing problem.

Another major technical challenge was to develop a method to scale discriminant measurement from Phase I in a way that no substantial changes in the existing adaptive logic would be required. The newly developed measurement was quite different in dynamic range and statistical properties than the original IFM measurement. Since Phase III tests were planned to evaluate measurement system differences, any adaptive logic change required by the different measurement systems could confound the evaluation, and was undesirable.

Analyses of the original IFM design rationale provided a solution. The original IFM measurement and adaptive logic design philosophy was based on assumptions of performance norms for experienced naval aviators which were derived from NATOPS standards. The score which represented one and two standard deviation performance could be expressed from these assumptions, and the adaptive logic algorithm was built on that premise. It was not possible to relate the assumptions of NATOPS standards to the more complex, discriminant measurement.

It was possible to empirically define criterion performance norms from measured performance data with the discriminant measurement, and to relate the old and discriminant measure distributions. When this was done, scoring of new measurement on each trial relative to the criterion performance distributions (expressed as z-scores) provided a method to equate the performance evaluation decisions by the adaptive logic for all measurement schemes.

In the process of working through this problem, it was noticed that the actual IFM performance score distributions were quite different than the assumed norms for the experienced naval aviators. One obviously simple and good way of improving measurement (of this type) would be to base measurement decisions on *actual* norms, rather than *assumed* norms. A third measurement subsystem based on actual IFM norms was designed and installed.

System engineering tests were conducted with two trainees with the result that real time measurement was achieved, and all measurement subsystems operated properly except discrim. The discriminant model measurement occasionally could misclassify very poor performance if that poor performance was on a negatively weighted measure. The problem was found to be caused

by poor performance exceeding the measurement space of the Phase I data. (The model was only valid within the measurement space of the data from which it was derived.) The problem was solved by establishing the 4-sigma boundaries for negatively weighted measures, and altering the real time discriminant scoring subsystem to first test for data boundaries. If the boundary was exceeded, the score was set to 2.7-sigma (a poor score). If the boundary was not exceeded for all negatively weighted measures, the discriminant function was computed. Subsequent tests were successful.

MEASUREMENT EVALUATION SUMMARY

The resulting three measurement subsystems were evaluated in Phase III by automatically training three matched groups of five civilian pilots each on the TRADEC/IFM-modified. Group I was trained using old IFM scoring. Group II was trained with discriminant model scoring. Group III was trained with normative IFM scoring. The raw results of time-to-train to the same performance criteria revealed that the discriminant model was far superior to either of the other measurement subsystems. However, the distributions of group matching variables were found to be unequal, biasing the results.

Removal of the significant sources of group bias (first score in the simulator and age) resulted in a 34-40% improvement in the time-to-train using both the discriminant model and the normative IFM model. Examination of typical trainee plots and a breakdown of the number of trials required to graduate from each maneuver suggested that the discriminant model provided more reliable performance feedback and it appeared to sense piloting technique.

The discriminant model was hypothesized to have greater growth potential than the normative IFM model because of the importance of student variables (such as age) and its ability to choose and properly weight significant student variables along with system performance variables and the measures of control activity.

The evaluation data also highlighted some potentially serious inefficiencies in the linear, single score adaptive logic design as it interacted with the syllabus and measurement system. These problems are discussed in Appendix F, and a study to explore more efficient logics is recommended.

The major conclusion and recommendation of the study was that the discriminant model should be applied to the problem of specifying measures for future flight training systems. In order to do that, empirical data must be collected at an operational training site to produce data for measure selection analyses. Additionally, the results of the measure selection analysis should be used to validate the effect of improved

NAVTRAEQUIPCEN 74-C-0063-1

measurement on training, similar to the methods employed in this study.

Remaining conclusions and recommendations can be found in summary outline form in Sections VI and VII.

SECTION II

MEASURE SELECTION METHOD

A combined analytic and empirical method was used to define measures for automated training of four instrument flight maneuvers. The method was based on the criteria that the final measure set should represent a comprehensive, yet minimum set of measures which (a) were sensitive to the skill change that occurred during training, (b) had performance prediction qualities, and (c) which tended to eliminate redundant forms of information. These criteria for measurement selection and the fundamental techniques and algorithms for selecting measures were developed and elaborated upon in earlier work (Vreuls, Obermayer, Goldstein and Lauber, 1973; Vreuls, Obermayer and Goldstein, 1974).

MEASURE SELECTION PROCESS SUMMARY

The measure selection process contained a series of related critical steps which began with an analysis of potential information needs for training. This first step involved the specification of performance measure candidates (candidates for empirical selection analyses) which in the judgment of the investigators (armed with data from earlier studies and sample analysis data) would contain information of importance to the adaptive logic which was to control training.

Next, the required raw data parameters, such as (but not limited to) vehicular state variables and their desired sampling rates were defined. Typically, raw data parameters were not in a form that was useful for automated measurement; however, error from desired values and transformations such as the average error contained the desired information. Potentially useful candidate measures were defined as transforms of parameters.

The conditions which define when measurement was to start and stop also were specified. It is emphasized that the specification of unambiguous rules to start measuring and to stop measuring can be underestimated; in practice, the construction of start/stop algorithms has been most challenging, and is a crucial part of performance measurement specification.

Having defined the measures and rules for obtaining measurement, the next step in the process required collection of empirical data during training to provide a battery of candidate measures for selection analyses. Computer measure selection analyses based on multivariate statistical models were used to reduce the measures to a final set according to each of the aforementioned criteria. The outcome of the analysis was interpreted by the investigators and merged with outer loop measures to form a final recommended set for each maneuver. Further computer analyses established the weighting coefficients

to use with each measure when combining the measures into a composite score for use by the adaptive logic.

The initial measure selection process included a combination of canonical correlation analysis and discriminant analysis. This method of selection proved more complex than fruitful and has been simplified to using discriminant analysis only: the canonical portions of the research have been deleted from this report.

APPARATUS

The test equipment was the Training Device Computer System (TRADEC) located at the Naval Training Equipment Center, Orlando, Florida. TRADEC was configured as a fixed-wing aircraft (F-4E). TRADEC hardware included an XDS Sigma-7 computer and associated peripherals, an aircraft cockpit mounted on top of a four degree-of-freedom motion platform (pitch, roll, yaw and heave), and a host of related equipment. The cockpit contained all of the controls and displays found in a jet fighter front seat, except that the radio navigation, communications and weapons systems were mocked up and non-functional. A digital computer program provided the basic flight simulation (cf Kapsis, et al, 1969; Erickson, et al, 1969).

The basic flight program was converted into a computer-controlled training device by an automated instrument flight maneuvers (IFM) program (cf Charles, Johnson and Swink, 1972). IFM automatically sequenced the trainee through a series of maneuvers and simulated flight conditions ordered from least to most difficult, as a function of measured trainee performance on the previous and antecedent trials. The performance measures and weighting coefficients for summing the various components of error into one composite score were derived during IFM system design from task analytic data. The measures were never formally tested.

As a part of a previous effort, IFM was modified to control a measure selection experiment and to produce raw data on magnetic tape for subsequent (non-real time) conversion into candidate measures to be used for measure selection analyses. A computer controlled speech synthesizer (COGNITRONICS) was used to brief participants on the task requirements for each trial, and issue corrective commentary when various vehicle states were out of selected tolerance bands based on NATOPS performance criteria. The IFM task scheduler was used to set the experimental conditions for the next trial as prescribed by the experimental design.

PARTICIPANTS

Twelve relatively low-time student and private pilots were used as trainees. They averaged 55 hours of flight time, 3.7 hours of prior instrument time and had a median age of 24 years.

All participants had some familiarity with instrument flight but were unskilled. All participants were light plane pilots and were unfamiliar with jet fighter responses.

TASK

Each participant was trained to fly four basic instrument flight maneuvers; (a) straight and level flight, (b) standard rate climbs and descents, (c) level turns and (d) climbing and descending turns. Aircraft weight and resultant center-of-gravity shift, and turbulence were varied.

Straight and level flight required the trainees to hold a heading of 360 degrees, altitude of 25,000 feet--Flight Level (FL) 250, airspeed of 350 knots for one-half of the trials and an airspeed of 280 for the remainder of the trials. Each trial lasted one minute.

Standard rate climbs and descents required the trainees to climb from FL 240 to FL 250, or to descend from FL 250 to FL 240 at a standard rate of 1,000 feet-per-minute while holding a 360 degree heading and 350 knots of airspeed. One-half of the trials were climbs; the other half were descents.

Level turns required constant bank (30 degree) turns from a heading of 360 degrees to a heading of 315 degrees or 045 degrees while holding FL 250 and 350 knots. One-half of the trials were left; the remainder were right.

Climbing and descending turns required a climb or descent for 1,000 feet at 1,000 feet-per-minute while turning through a 90 degree heading change and holding airspeed at 280 knots; the initial climb or descending turn was followed by a reversal of turn direction and altitude rate, and subsequent return to the starting heading and altitude. One-half of the trials were left, descending turns starting at FL 250, followed by a right climbing turn back to FL 250 and heading 360 degrees. The remaining trials were right, climbing turns starting at FL 240, followed by descending, left turns back to FL 240 and heading of 360 degrees.

Two task stressors were used, turbulent air and aircraft weight and center of gravity. The turbulent air was produced in the flight program by a random number generator. When used, its intensity was set to a "light turbulence" level as defined by the IFM program. The aircraft weight was either light or heavy. The light aircraft carried 2500 pounds of fuel, had a gross weight of 33,600 pounds and center-of-gravity at 29.0 percent mean aerodynamic chord. The heavy aircraft carried 12,896 pounds of fuel, had a gross weight of 43,996 pounds and a center-of-gravity at 30.2 percent mean aerodynamic chord. The weight increase and aft center-of-gravity shift reduced the longitudinal axis short-period damping coefficient, which decreased the simulator pitch axis stability, making it more difficult to control. Task stressors were not changed during a trial.

PROCEDURE

Each participant was given a familiarization flight to learn the experimental procedures and the simulator. No participant started the experiment until they had the simulator under control, according to the judgment of the Test Director, who was a Commercial Pilot with Instrument, Helicopter, Sailplane, Multi-Engine, Land and Seaplane ratings.

IFM trimmed the simulator for straight and level flight at the initial heading, altitude and airspeed prior to the beginning of each trial. The trainee was instructed on the conditions of the run by the COGNITRONICS and told to take control. In addition, a card diagramming each maneuver for each trial was placed in the cockpit for reference. The trial and data collection were trainee initiated by placing the speed brake switch forward. Speed brake aerodynamic effects were locked-out of the simulation software.

EXPERIMENTAL DESIGN

The participants were trained on the four basic instrument flight maneuvers for 18, one-hour sessions. A total period of 19 weeks was required to collect the data. Six trials of each maneuver were flown during each training session. Each successive odd and even numbered training session was pooled into one unit called a training "day"; thus, sessions 1 and 2 became Day 1, sessions 3 and 4 became Day 2, etc. This pooling resulted in 144 possible observations for each maneuver on a given training day (12 participants by 6 trials by 2 sessions). The design is shown in table 1.

Each participant received exactly the same order of experimental trials on each day. Thus, maneuver one always was flown first and maneuver four always was flown last. This fixed order permitted the study of measures for each maneuver under identical antecedent conditions (and subsequent order effects) across training days.

On Days 1, 3, 5, and 7 the trials were flown with a light aircraft (forward C.G.) and no turbulence. A heavy aircraft (aft C.G.) was presented on Days 2, 4, and 6 without turbulence. Light turbulence was presented on Day 8 with a light aircraft. Day 9 consisted of a heavy aircraft and light turbulence.

It was assumed that after 14, one hour training sessions (the conclusion of Day 7), the trainees would be relatively proficient on the basic maneuvers. Therefore, a comparison of performance differences between Day 1 and Day 7 should reveal measures which were sensitive to the skill change that occurred and those measures which had performance prediction qualities without task stressors in operation. A similar comparison of Day 2 versus Day 6 should reveal those measures which are sensitive to training when flying with an aft C.G.;

TABLE 1. EXPERIMENTAL DESIGN

		G1T1				G2T1			G1T2	G2T2
		DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	DAY8	DAY9
M1	P1	*	*	*	*	*	*	*	*	*
	P2	*	*	*						
	.	*								
	P12									
M2	P1	*								
	P2	*								
	.	*								
	P12									
M3	P1	*								
	P2	*								
	.	*								
	P12									
M4	P1									
	P2	*								
	.	*								
	P12	*								

Legend: M=Maneuvers:

M1 = Straight and Level
 M2 = Standard Rate Climbs and Descents
 M3 = Level Turns
 M4 = Climbing and Descending Turns

P=Participants

G=Center of Gravity:

G1 = Light Aircraft, Fore cg.
 G2 = Heavy Aircraft, Aft cg.

T=Turbulence:

T1 = Smooth Air
 T2 = Light Turbulence

DAY=Two successive one-hour training sessions.

* Twelve trials were administered on each maneuvers, each day.

the measure set was not expected to be exactly the same as with a forward C.G.

A comparison of Day 7 versus Day 8 performance was anticipated to reveal the differences in measure set composition caused by the addition of turbulence as a task stressor. Day 7 versus Day 9 performance would provide evidence of the measure set differences caused by the addition of both aft C.G. and turbulence as task stressors.

The development of this experimental design presented challenges which required compromise between theoretical issues and practical constraints. Our biases in attacking these compromises were poignantly expressed by Cooley and Lohnes (1971, p.v.), drawing reference to Tukey (1962):

Tukey argued that there have to be people in the various sciences who concentrate much of their attention on methods of analyzing data and of interpreting the results of statistical analysis. These have to be people who are more interested in the sciences than in mathematics, who are temperamentally able to 'seek for scope and usefulness rather than security,' and who are 'willing to err moderately often in order that inadequate evidence shall more often suggest the right answer.' They have to use scientific judgement more than they use mathematical judgement, but not the former to the exclusion of the latter. Especially as they break into new fields of sciencing, they must be more interested in 'indication procedures' than in 'conclusion procedures' (or in conclusions that must be considered statistically weaker).

It was recognized that there was a possibility of confounding the effects of further training beyond Day 7 with the effects of task stressors in the experimental design. However, the design was the only practical one because of the length of time required to collect data (19 weeks). A full factorial design with all conditions presented on each training day would have reduced the number of observations of a given condition to the extent that multivariate measure selection techniques would not have been possible, because increasing the number of participants and data collection time was not possible within the scope of the current effort. It was later found that weighting coefficients for turbulence tasks could not be accurately determined and turbulence was dropped as a task stressor in the final task syllabus for evaluation purposes.

The purpose of the study was to select a minimum number of measures from a larger, candidate measure battery. Earlier work suggested that multivariate methods offered a good avenue for problem solution. Several mathematical issues were brought about by our desire to explore multivariate models as a basis for measure selection algorithms.

One issue was the number of observations, or samples required in each experimental group. For measure selection purposes, Lane (1971) concluded that five-to-seven times as many participants (samples) as initial measures in a battery are required for multiple regression analysis to adequately address shrinkage and overfit. Extrapolating Lane's criterion to our current problem revealed that 112 to 144 participants would be required with 16 initial candidate measures in the test battery. Since it took 19 weeks to collect data from 12 participants, it would have taken 228 weeks to collect data from 144 participants. Clearly, this was not possible.

It was possible to form 144 observations for each day of training by pooling data from 12 repeated trials of each of 12 participants. The consequences of pooling data in this way to produce a sufficient number of scores for proper operation of the multivariate models were unclear at the onset of the study. A review of the literature and informal consultation with several statisticians resulted only in the conclusion that the problem was a researchable issue.

Classical multivariate techniques have been used in personnel selection and classification for years, and are well developed for that purpose. Most of the literature addresses the classification problem, which typically asks questions about the probability of group membership of an individual with certain measured traits. These classification techniques require familiar assumptions of independent sampling of various populations to achieve assumed multivariate normal distributions and equality of dispersions.

Our research problem, however, was not to assess the probability of group membership, but to find a method that would display measure changes for given individuals as a consequence of their training.

The best tool for finding measures appeared to be the multiple discriminant model which is well defined in the following excerpt from Cooley and Lohnes (1971, p. 243):

The discriminant model may be interpreted as a special type of factor analysis that extracts orthogonal factors of the measurement battery for the specific task of displaying and capitalizing upon differences among the criterion groups. The model derives the components which best separate the groups of a taxonomy in the measurement space. It makes no difference to the formal logic of the model whether the samples of several populations are viewed as the dependent, criterion variable and the discriminant functions are viewed as the best prediction functions of the independent, predictor vector variable defining the measurement space, or if the groups are viewed as the independent treatment variable and the discriminant functions are seen as the most predictable

functions of the dependent vector variable. The taxonomic variable is more likely to be the criterion variable in survey science, whereas it is almost certain to be the independent treatment variable in experimental research.

The latter definitions of variables appeared to fit the current research problem; groups may be considered to be the independent variable. The different treatment groups (or days) represent a continuum from early to late training. Selection of specific comparisons constituted samples from the continuum.

The experimental design shown in table 1 reveals that group membership was fixed by assignment of the same people to each group (day). We could not increase the number of participants to form two independent groups because data collection time would have doubled. Neither could we decrease the number of maneuvers (in order to increase the number of participants within the same data collection time frame) because measurement information was needed for each maneuver and each task stressor.

Assignment of the same people to each group and repeated observations in each group may violate the assumption of independent sampling; however, these violations were necessary, and may not be severe. When assumptions are obviously violated, Winter (1974) indicated that the linear discrimination model simply becomes an empirical procedure, which although it may not be optimum, may be satisfactory from a practical viewpoint. Also, to counteract some of the effects of this violation on the data, removal of these components of variance was done before discriminant analysis was performed. There can be no doubt that the discriminant model will find and highlight the measurement components that best display the differences between groups, as it was used herein as part of an empirical procedure. The procedure should be validated in future efforts.

MEASUREMENT

RAW DATA. Eighteen pilot/system performance parameters shown in Appendix A, were recorded on magnetic tape at a rate of five times-per-second in real time from the beginning to the end of training. The raw data were checked and packed onto 16 reels of 2400 foot, 9 track magnetic tape in binary format. These data were processed after data collection was complete. Measures were created from the raw data by computer programs designed to execute the approach to measurement which had been previously developed for NAVTRAEQUIPCEN by the authors.

MEASUREMENT APPROACH. A description framework has been established which relates system performance and human behavior to segments of maneuvers constituting a training mission. This descriptive structure has been called a measurement model. The model permits the measurement of a variety of tasks and performance dimensions in order to describe unique as well as common

aspects of maneuvers. To accomplish this, the model defines each measure in terms of the following six determinants (which are summarized in the paragraph below): (a) A maneuver segment; (b) A parameter; (c) A sampling rate; (d) A desired value if required; (e) A tolerance value if required, and (f) A transformation.

A segment is any portion of a maneuver for which desired student behavior or system performance follows a lawful relationship from beginning to end, and for which the beginning and end can be unambiguously defined. The measurement start and stop conditions define a segment. A parameter is any quantitative index of (a) vehicle states in any reference plane, (b) personnel physiological states, (c) control device states, or (d) discrete events. A sampling rate is the temporal frequency at which the parameter is examined. Frequently parameters have no utility unless compared to a desired value or a tolerance to derive an error score. Finally, a transformation is any mathematical treatment of the parameter, to include measures of central tendency, variability, scalar values, Fourier transforms, pilot/system transfer functions, etc.

The reader is urged to take careful note of the definition of a measure used throughout this report; a measure is the end result of the measure production process, which starts with a raw data parameter and ends with a specific transformation of that parameter.

Current measure producing computer program functions for defining measurement start/stop conditions and logically combining start/stop expressions are shown in Appendix A. Common measurement transformations available in the measurement programs are shown also in Appendix A.

CANDIDATE MEASURES. The raw data were processed by the measurement software to produce candidate measures for measure selection analyses. Candidate measures for each maneuver are shown in tables 2 - 6. The tables indicate, from left to right, the parameter variable names in the simulation software, the desired value(s), the transform names in the measurement software and the measure abbreviation used throughout the report. Segmentation rules are noted for each maneuver.

Maneuver 4 was subdivided into three segments, numbered 2, 3, and 4. Segment 2, Initial Climb or Descent, started at the beginning of the climbing or descending turn and continued until a change in altitude had exceeded 1,000 feet, and heading had changed from the initial value by more than 90 degrees. Segment 3, Climb or Dive and Turn Reversal, started at the end of Segment 2, and continued until altitude had returned within 1,000 feet of the initial altitude. Segment 4, Final Climb or Descent, started when altitude was within 1,000 feet of the initial altitude and heading was within 90 degrees of the initial value, and ended at the end of the maneuver as defined by the IFM program.

TABLE 2. CANDIDATE MEASURES FOR MANEUVER 1, STRAIGHT AND LEVEL

MEAS. NO.	PARA-METER	DESIRED VALUE	TRANS-FORM	ABBREVIATION IN ANALYSIS	GLOSSARY
1	ELVS	0	RNG	ELRG	ELEVATOR STICK RANGE
2			FLTR	ELF1	CROSSOVER POWER
3			AAE	ELF2	AVERAGE DISPLACEMENT
4	AILS	0	RNG	AIRG	RANGE
5			FLTR	AIF1	CROSSOVER POWER
6			AAE	AIF2	AVERAGE DISPLACEMENT
7	PED	0	RNG	PDRG	RANGE
8			FLTR	PDF1	CROSSOVER POWER
9			AAE	PDF2	AVERAGE DISPLACEMENT
10	ALPH	0	RNG	ALRG	RANGE
11			SDEV	ALSD	STANDARD DEVIATION
12	PTCH	0	RNG	PTRG	RANGE
13			SDEV	PTSD	STANDARD DEVIATION
14	ROLL	0	AAE	RDAA	ABSOLUTE AVERAGE ERROR
15			RMS	RORM	ROOT-MEAN-SQUARED ERROR
16	HEAD	360	RMS	PSRM	ROOT-MEAN-SQUARED ERROR
17			RNG	PSRG	RANGE
18	ALT	25000	AAE	HAA	AVERAGE ABSOLUTE ERROR
19			RNG	HRG	RANGE
20	HDOT	0	AAE	HDA A	AVERAGE ABSOLUTE ERROR
21			RNG	HDRG	RANGE
22	A/S	350/280 ¹	AAE	ASAA	AVERAGE ABSOLUTE ERROR
23			RNG	ASRG	RANGE
Segmentation Rules: Start Meas. at the Beginning of the Trial (Speed Balance In) Stop Meas. at the End of the Trial (One Minute of Elapsed Time).					

¹One-half of the trials were at 350-knots IAS, the other half at 280-knots.

TABLE 3. CANDIDATE MEASURES FOR MANEUVER 2, CLIMBS AND DESCENTS

MEAS. NO.	PARA-METER	DESIRED VALUE	TRANS-FORM	ABBREVIATION IN ANALYSIS	GLOSSARY
1	ELVS	0	FLTR	ELF1	ELEVATOR STICK
2			AAE	ELF2	CROSSOVER POWER
3	ALPH	0	RNG	ALRG	AVERAGE DISPLACEMENT RANGE
4			SDEV	ALSD	STANDARD DEVIATION
5	PTCH	0	SDEV	PTSD	STANDARD DEVIATION
6	HDOT	1000 ¹	AAE	HDAA	AVERAGE ABSOLUTE ERROR
7	AILS	0	FLTR	AIF1	CROSSOVER POWER
8			AAE	AIF2	AVERAGE DISPLACEMENT
9	ROLL	0	AAE	RDA A	AVERAGE ABSOLUTE ERROR
10			RMS	RORM	ROOT-MEAN-SQUARED ERROR
11	PED	0	FLTR	PDF1	CROSSOVER POWER
12			AAE	PDF2	AVERAGE DISPLACEMENT
13	HEAD	360	AAE	PSAA	ABSOLUTE AVERAGE ERROR
14			RMS	PSRM	ROOT-MEAN-SQUARED ERROR
15	TURN	0	RMS	TURN	ROOT-MEAN-SQUARED ERROR
16			AAE	TUAA	ABSOLUTE AVERAGE ERROR
17	BETA	0	RMS	BERM	ROOT-MEAN-SQUARED ERROR
18	A/S	350	AAE	ASAA	ABSOLUTE AVERAGE ERROR
19	THRR	0	RNG	THRG	ABSOLUTE AVERAGE ERROR RANGE
Segmentation Rules: Start Meas. at Beginning of Trial (Speed Brake In). Stop Meas. at End of Trial (1,000 Foot Altitude Change).					

¹ One-half of the trials were climbs at 1000 FPM, the other half were descents at-1000FPM.

TABLE 4. CANDIDATE MEASURES FOR MANEUVER 3, LEVEL TURNS

MEAS. NO.	PARA-METER	DESIRED VALUE	ABBREVIATION IN TRANS-FORM ANALYSIS	GLOSSARY
1	ELVS	0	FLTR	ELEVATOR STICK
2			AAE	CROSSOVER POWER
3	ALPH	0	RNG	AVERAGE DISPLACEMENT
4			SDEV	RANGE
5	PTCH	0	SDEV	STANDARD DEVIATION
6	AILS		FLTR	STANDARD DEVIATION
7			AAE	CROSSOVER POWER
8	ROLL	30 ¹	AAE	AVERAGE DISPLACEMENT
9			AAE	AVERAGE ABSOLUTE ERROR
10	PED	0	RMS	ROOT-MEAN-SQUARED ERROR
11			FLTR	CROSSOVER POWER
12	BETA	0	AAE	AVERAGE DISPLACEMENT
13			RNG	RANGE
14	A/S	350	RMS	ROOT-MEAN-SQUARED ERROR
15			AAE	AVERAGE ABSOLUTE ERROR
16	ALT	25000	RMS	ROOT-MEAN-SQUARED ERROR
17	THRR	0	AAE	AVERAGE ABSOLUTE ERROR
			RNG	RANGE
Segmentation Rules: Start Meas. at Beginning of Trial (Speed Brake In). Stop Meas. at End of Trial (45° Heading Change).				

¹ One-half of the trials were right (+) 30 degree bank turns, the other half were left (-) 30 degree bank turns.

TABLE 5. CANDIDATE MEASURES FOR MANEUVER 4, CLIMBING AND DESCENDING TURNS

MEAS. NO.	PARA-METER	DESIRED VALUE	TRANS-FORM	ABBREVIATION IN ANALYSIS	GLOSSARY
1	ELVS	0	FLTR	ELF1	ELEVATOR STICK
2			AAE	ELF2	CROSSOVER POWER
3	ALPH	0	RNG	ALRG	AVERAGE DISPLACEMENT
4	HDOT	1000 ¹	AAE	HDAA	ANGLE OF ATTACK
5	THRR	0	RNG	THRG	ALTITUDE RATE
6	A/S	280	AAE	ASAA	RIGHT THROTTLE
7	AILS	0	FLTR	AIF1	AIRSPEED
8			AAE	AIF2	AILERON STICK
9	BETA	0	RMS	BERM	SIDE SLIP
10	ROLL	30 ¹	AAE	ROAA	ROLL ATTITUDE
11	PED	0	FLTR	PDF1	RUDDER PEDAL
12			AAE	PDF2	CROSSOVER POWER
13	HEAD	90 ¹	AFIN	HDAF	AVERAGE DISPLACEMENT
14	TIME	0	ELT	TIME	ABSOLUTE FINAL VALUE
Segmentation Rules for Initial Climb/Dive Turn:					ELAPSED TIME
Start Meas. at Beginning of Trial (Speed Brake In).					
Stop Meas. when (Altitude change was greater than 1000 feet).AND. (Heading was less than 270 degrees or greater than 90 degrees).					
Segmentation Rules for Final Climb/Dive Turn:					
Start Meas. when (Altitude change was within 1000 feet of Initial).AND. (Heading was greater than 270 degrees or less than 90 degrees).					
Stop Meas. when Altitude Returns to Initial Value.					

¹ One-half of the trials started with right, climbing turns, then reversed to left descending turns. The desired values were changed appropriately as a function of maneuver segmentation.

TABLE 6. CANDIDATE MEASURES FOR MANEUVER 4, CLIMB OR DIVE AND TURN REVERSAL

MEAS. NO.	PARA-METER	DESIRED VALUE	TRANS-FORM	ABBREVIATION IN ANALYSIS	GLOSSARY
1	ELVS	0	FLTR	ELF1	ELEVATOR STICK
2			AAE	ELF2	CROSSOVER POWER
3	ALPH	0	RNG	ALRG	AVERAGE DISPLACEMENT
4	HDOT	1000 ¹	AFIN	HDAF	RANGE
5	AILS	0	FLTR	AIF1	ABSOLUTE FINAL VALUE
6			AAE	AIF2	CROSSOVER POWER
7	BETA	0	RNG	BERG	AVERAGE DISPLACEMENT
8	ROLL	30 ¹	AFIN	ROAF	RANGE
9	TIME	0	ELT	TIME	ABSOLUTE FINAL VALUE
10	PED	0	FLTR	PDF1	ELAPSED TIME
11			AAE	PDF2	CROSSOVER POWER
					AVERAGE DISPLACEMENT

Segmentation Rules: Start Meas. when (Altitude change was greater than 1000 feet from Initial). AND. (Heading was less than 270 degrees or greater than 90 degrees).
Stop Meas. when (Altitude was greater than FL24 and less than FL25).

¹One-half of the trials started with right, climbing turns, then reversed to left descending turns. The values changed appropriately as a function of measurement segment.

MEASURE SELECTION ANALYSES

Measure selection analyses were performed by univariate and multivariate techniques. The results were interpreted by the investigators and merged into a composite measure set for each maneuver. Final discriminant analyses were performed to determine the relative weights of the recommended measures.

UNIVARIATE SELECTION. Considering each measure independent of all other measures, the average value of each measure on a given day was compared to the average value of that measure on the criterion day. A t-test was used to determine statistically significant differences. The means on Days 1, 3, and 5 were tested against Day 7 for performance changes during training without turbulence and with a light aircraft. The means for Days 2 and 4 were tested against Day 6 for performance changes during training with a heavy aircraft. Day 8 means were tested against Day 7 means to find the significant changes caused by the addition of light turbulence. Day 9 means were compared to Day 7 means to determine the measure set changes brought about by the addition of both light turbulence and a heavy aircraft.

DISCRIM SELECT. Computer programs have been generated to select measures through multiple discriminant analyses (cf Cooley and Lohnes, 1971). These analyses assume that a battery of measures have been taken for each of a number of groups of participants. The primary purpose of DISCRIM SELECT is to isolate the measures that best discriminate between groups. For example, a pair of groups may consist of experienced and inexperienced participants; the procedure adopted discards measures that do not contribute to such discriminations when all measures are considered together as a set.

A data editing and sorting routine was added to the initial part of DISCRIM SELECT in order to facilitate the components of variance removal programs. (See figure 1.) The components of variance programs required that the data be sorted according to subject, trial, day and maneuver and that all erroneous data be predetermined so that matching cells can be formed across days used in the analysis. For example, if the data for Subject 1, Day 1, and Maneuver 3 were erroneous in an analysis of Day 1 paired with Day 7, neither Day 1 or Day 7 data for Subject 1 and Maneuver 3 would be present in the analysis.

Two programs were designed to remove from the data the effects of observing the same subjects in all conditions and the effects of observing the same subject twelve times in each condition. Both programs subtracted the components of variance from each data point. RMEAS subtracted the effect of observing the same subject on both days (as suggested by Schori, 1972):

$$X_{mikL} = X_{mikL} - \left(\frac{1}{K} \sum_{k=1}^K X_{mikL} - \frac{1}{NKR} \sum_{L=1}^N \sum_{K=1}^K \sum_{L=1}^R X_{mikL} \right)$$

where: m - Variables M - No. of variables
 i - Subjects N - No. of subjects
 k - Groups K - No. of groups
 L - Observations/day R - No. of observations/day

REPM subtracted the effect of observing the same subject more than once in a day using:

$$X_{mikL} = X_{mikL} - \left(\frac{1}{R} \sum_{L=1}^R X_{mikL} - \frac{1}{NR} \sum_{L=1}^N \sum_{L=1}^R X_{mikL} \right)$$

Both of these operations were performed before any other statistical analysis.

Many measures were transforms of closely related parameters. Highly correlated measures were eliminated in order to reduce redundant information, and to avoid computation problems which were experienced with trial data when intercorrelations greater than $r=.95$ existed in the candidate measure sets. The criterion for dropping one member of a highly correlated pair was established by tests with $r=.95$, $r=.90$ and $r=.80$; $r=.90$ was selected because it appeared to eliminate obvious redundancies, yet left a reasonable number of measures for subsequent analyses. Since measure transforms were ordered, generally, from easiest to most difficult to compute in the candidate lists, the procedure was adopted to drop the most difficult to compute an intercorrelated pair for a given maneuver and analytic comparison.

DISCRIM SELECT iteratively discarded measures until a minimum set of measures resulted. The iterative process stopped when either one of two criteria was met, (a) the total number of remaining measures was less than the minimum number of Factors required to describe the variance as determined by a Principal Components Analysis, or (b) discarding another measure would have reduced the overall discrimination to an unacceptable level.

Two tolerances associated with the above criteria had to be specified by the investigators, (a) the minimum percent variance to be accounted for by any Factor, and (b) the minimum measurement communality. Communality was the amount of variance a particular measure contributes to all discriminant functions.

The tolerances were set by trial analyses with maneuver one data. It was found that between 90 and 95 percent of the original variance was retained when the minimum variance for any Factor was set at 7 percent; this tolerance set the minimum possible measure set size to equal the minimum number of "significant" Factors. Trial analyses also revealed that in most cases measures which exhibited communalities less than .300 were non-significant contributors to the discriminant function, as shown by the Multivariate Analysis of Variance (included in DISCRIM SELECT software). Minimum communality was set at .300.

The flow diagram for DISCRIM SELECT is shown in figure 1; each block is described in the following:

1. Read tolerances and measure tables. The level of correlation for the initial removal of equivalent measures, and the labels for each measure, were read from punched cards at the beginning of the program. The two additional criteria were read for DISCRIM SELECT, (a) the minimum variance and (b) the minimum communality.
2. List initial measure set. The initial measure set was listed by number and name of each measure.
3. Sort all data in the selected groups according to subject, day, maneuver and trial. Match cells when rejecting erroneous data.
4. Perform removal of components of variance to correct for repeated observations.
5. Combine data from two selected groups. Measures from one time in training were to be compared to the same measures taken at another time in training. The measures from each training day, or each group, were brought together into a common data file so that the same types of measures could be compared observation by observation.
6. Generate correlation matrix. Each measure was correlated with every other measure to form an intercorrelation matrix.
7. Remove highly correlating measures. One member of a pair of measures was removed from further analysis when the correlation coefficient in the matrix exceeded 0.90. The candidate measures were ordered, generally, from least to most difficult to compute. The more difficult to compute transform of a measure-pair was dropped. No measure was removed for reason of high correlation if the high correlation coefficient occurred between two measures taken at different points in training.
8. List measures kept and dropped. The measures were again listed in two columns, one column for those kept for further analysis, and the other column those which were removed from the analysis.

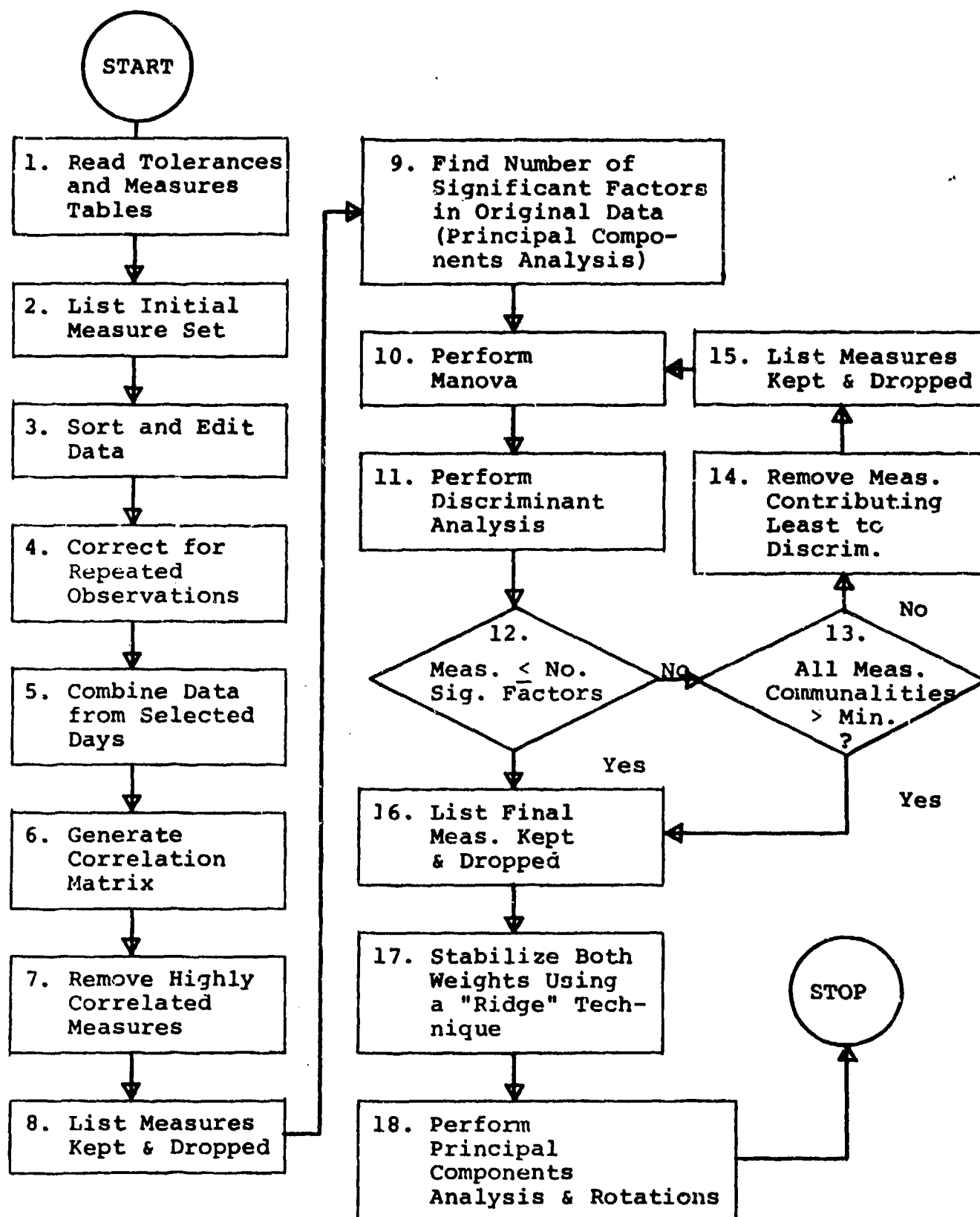


Figure 1. DISCRIM SELECT Functional Flow.

9. Perform principal components analysis and rotations, which produced the following outputs:

- a. Factor structure for each group (Day) of the comparison provided evidence of performance dimensions.
- b. The percent variance explained by each Factor, degrees-of-freedom and CHI SQUARE aided in the assessment significant factors. The percent of variance explained by each factor was used (in Step 12) to establish the minimum number of measures.
- c. VARIMAX rotations which were used to present the principal dimensions of performance changes.

10. Perform multivariate analysis of variance (MANOVA), producing the following output for use by the discriminant analysis:

- a. Means and standard deviations by group.
- b. A test for equality of dispersions.
- c. Univariate F-ratios for each measure used to establish reasonable grounds for the commonality measure rejection criterion (Step 11).
- d. Multivariate test of significant, Wilks' Lambda and F-ratios.

11. Perform multiple discriminant analysis (DISCRIM), producing the following information:

- a. Multivariate test of significance, Wilks' Lambda and F-ratios (a check on 10. d).
- b. CHI-SQUARE with successive roots removed provided evidence of the statistical significance of the discriminant function (since only one discriminant function was generated).
- c. Measure coefficient vectors, the weights to combine the measures into the discriminant function.
- d. Communalities, the proportion of variance (associated with each measure) extracted by all discriminant functions; as noted previously, communality was the basis of removing measures from the set.
- e. Group centroids in discriminant space revealed the group mean position on the discriminant function.

12. If the number of measures remaining was less than or equal to the number of significant factors, iterative measure elimination ceased and the program branched to Block 16. If the number of measures was greater than the number of significant factors, the program continued to the next test.

13. All remaining measure communalities were tested against the experimenter specified minimum communality (set at 0.30 in this study). If no communalities were less than criterion, the program terminated through Block 16. If there were remaining communalities less than criterion, iterative measure elimination continued.

14. The measure with the least communality was found and eliminated from the set and correlation matrix.

15. The measures kept and dropped were listed, and the analysis was recomputed starting at Block 10.

16. The final measures retained and those dropped in order of elimination were listed.

17. Perform "ridge" analysis by iteratively adding bias to W matrix and reperforming DISCRIM.

18. A final principal components analysis (and rotations) was performed to show the ending factor structure.

The resulting set was examined to insure that all vehicular outer loops which represented task instructions (such as hold heading, airspeed and altitude) were represented. If outer loop measures were dropped during iterative analyses, they were added back into the recommended set.

Finally, DISCRIM SELECT was modified to perform an analysis on only the recommended measure set in order to assure that a significant discriminant function was retained, and to compute the weights assigned to each measure of the final set for combining data into a single score, the discriminant function, for each maneuver, setment and day comparison group.

In order to explore the reliability/stability of the discriminant model DISCRIM SELECT was also modified to add a bias (in 0.1 increments) to the diagonal of the W matrix and then reperform DISCRIM under operator control after the recommended measure set was determined. This is referred to in the flow chart and was similar to "Ridge" regression analysis. (Hoerl and Kennard, 1970.)

SECTION III

MEASURE SELECTION RESULTS AND DISCUSSION

Candidate measure sets were created differently for each of the instrument flight maneuvers to reflect different dimensions of control and different criteria of performance. Results were presented for each maneuver. Within each maneuver there were four day-comparisons, which represented changes in task complexity. The first two day-comparisons sought the measures which would reveal performance changes from initial to final training with no stress -- (a) light aircraft, forward C.G. and no turbulence -- (Day 1 vs Day 7, and (b) with heavy aircraft, aft C.G. and no turbulence -- (Day 2 vs Day 6). The third and fourth day-comparisons sought the measure set changes required by the addition of (a) turbulence only (Day 7 vs Day 8) and (b) turbulence combined with a heavy aircraft and aft C.G. (Day 7 vs Day 9).

Summary data are presented in this section in accordance with four steps in the measure selection process, (a) means and t-tests, (b) removal of equivalent measures, (c) multiple discriminant selection analyses (DISCRIM SELECT), and (d) the recommended measures and weighting coefficients for summing the set into one composite score for each maneuver.

MEANS AND t-TESTS

The average values of each measure for every maneuver and segment are presented in Appendix B for each training day. Almost all of the measures exhibited a reduction in error as a function of training day, which lent face validity to the training sensitivity of the initial candidate measure set. Each of the day-comparisons were tested for significant differences by t-tests. Those measures which were significantly different for each of the comparisons were selected as contributors to the training sensitive measure set.

Results were summarized in table 7. Generally, more measures were selected for less complex tasks (Maneuvers 1 and 2) than for the more complex Maneuvers 3 and 4. Since Maneuvers 1 and 2 were felt to be less demanding than Maneuvers 3 and 4, these data suggested that either a sufficient set for the more complex tasks was not constructed, or there were more redundant forms of measurement in the first two maneuvers.

EQUIVALENT MEASURES

The number of equivalent measures for each maneuver and day-comparisons are shown in table 8. Measures which intercorrelated greater than $r=.90$ were considered to be equivalent in this and subsequent analyses, and therefore could be substituted for one another. It was noted that more equivalent measures were found in Maneuvers 1 and 2 than in the remaining maneuvers.

TABLE 7. NUMBER OF MEASURES SELECTED BY t-TESTS

MANEUVER	(NCM) ¹	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9	ROW MEAN ³
1. St. & Level	(23)	21	20	13	14	17
2. Climbs/Dives	(19)	18	17	13	13	15
3. Level Turns	(17)	13	11	9	10	11
4-2. Initial CDT ²	(14)	6	6	6	6	6
4-3. CDT Reversal	(11)	3	7	5	6	5
4-4. Final CDT	(14)	11	8	6	9	9
Column Mean ³		12	12	9	10	10
		NO STRESS	AFT C.G.	ADD TURBULENCE	ADD AFT TURBULENCE	C.G.

TABLE 8. NUMBER OF EQUIVALENT MEASURES⁴

MANEUVER	(NCM) ¹	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9	ROW MEAN ²
1. St. & Level	(23)	10	8	8	8	9
2. Climbs/Dives	(19)	7	6	7	8	7
3. Level Turns	(17)	3	2	3	4	3
4-2. Initial CDT ³	(14)	5	1	1	4	3
4-3. CDT Reversal	(11)	4	3	3	3	3
4-4. Final CDT	(14)	2	0	1	2	1
Column Mean ²		5	3	4	5	4
		NO STRESS	AFT C.G.	ADD TURBULENCE	ADD AFT TURBULENCE	C.G.

¹NCM = Number of Candidate Measures²Means were rounded to the nearest whole number³CDT = Climbing and Diving Turns⁴A given measure may be equivalent to more than one measure

The composition of the equivalent measure forms is shown in Appendix C. Maneuvers 1 and 2 produced large chains of equivalent measures. In particular, pitch axis control range (ELRG) resulting angle-of-attack and pitch attitude (ALRG, ALSD, PTRG and PTSD) were highly correlated throughout Maneuver 1. Another cluster of redundant forms appeared for altitude and altitude rate (HRG, HDAA and HDRG). Roll absolute error and rms (ROAA and RORM) were equivalent for Maneuvers 1 and 2. Aileron and pedal displacement (AIF2 and PDF2) and resulting sideslip (BERM) were equivalent during Maneuver 2, climbs and dives.

The climbing and diving turn reversal segment was quite interesting because elevator stick, aileron stick and pedal crossover power (ELF1, AIF1 and PDF1) were equivalent to each other and to the final roll attitude value achieved (ROAF). Aileron and pedal displacement (AIF2 and PDF2) were equivalent only during training under no stress conditions, Day 1 vs 7.

The equivalent measures analysis served as a valuable first step filter to eliminate unnecessary measurement. The number of measures remaining after removal of equivalent forms is shown in table 9. The following multivariate measure selection analyses received a maximum of 15 measures to operate upon. Given 144 observations for each measure, the worst case (15 measures) for multivariate analyses produced 9.5 observations-per-measure, which was within the limits set by Lane (1971), making the assumption that observations x subjects were equivalent to subjects alone after the "components" of variance were removed.

TABLE 9. NUMBER OF MEASURES REMAINING FOR MULTIVARIATE ANALYSES

MANEUVER	(NCM) ¹	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9	ROW MEAN ²
1. St. & Level	(23)	15	15	15	15	15
2. Climbs/Dives	(19)	12	13	12	11	12
3. Level Turns	(17)	13	14	13	13	13
4-2. Initial CDT ³	(14)	9	13	13	10	11
4-3. CDT Reversal	(11)	7	8	8	7	8
4-4. Final CDT	(14)	12	14	13	12	13
Column Mean ²		11	13	12	11	12
		NO STRESS	AFT C.G.	ADD TURBULENCE	ADD AFT C.G. & TURBULENCE	

¹NCM = Number of Candidate Measures

²Means were rounded to nearest whole number

³CDT = Climbing and Diving Turns

DISCRIM SELECT

The components of variance removal routines increased the symmetry of the raw data while preserving the group centroids. With the improved dispersions of the data and the effects of repeated observations removed, the basic assumptions of multivariate discriminant analysis were met. The relationship of the measures between days then could be determined more accurately.

The multiple discriminant analysis iteratively reduced the measure sets, removing measures which contributed little to the discriminant function. Measures with low communalities (less than .30) were dropped, one at a time, and the process was repeated. Iteration continued until there were no measures left with low communalities, or the number of measures were equal to the number of factors which accounted for more than seven percent of the variance in the initial measure set.

After elimination of redundant measures, DISCRIM SELECT further reduced the candidate measures to an overall average of six measures for each comparison-day and maneuver. The data in table 10 illustrated that slightly more measures were required during Maneuvers 1 and 2 than during the remaining comparisons. Only three-to-five measures were sufficient to describe performance changes due to adding turbulence (Day 7 vs 8) and turbulence combined with a heavy aircraft (Day 7 vs 9) as task stressors.

TABLE 10. NUMBER OF MEASURES IN EACH MINIMUM DISCRIMINATING SET

MANEUVER	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9	ROW MEAN ¹
1. St. & Level	12	11	3	5	8
2. Climbs/Dives	12	9	5	4	8
3. Level Turns	9	6	5	4	6
4-2. Initial CDT ²	6	5	5	4	5
4-3. CDT Reversal	5	5	4	5	5
4-4. Final CDT	5	4	5	4	5
Column Mean ¹	8	7	5	4	6
	NO STRESS	AFT C.G.	ADD TURBULENCE	ADD AFT C.G. & TURBULENCE	

¹Means were rounded to nearest whole number

²CDT = Climbing and Diving Turns

NAVTRAEQUIPCEN 74-C-0063-1

The composition of the minimum discriminant set changed as a function of training (Day 1 vs 7 and Day 2 vs 6 taken together) and as a function of task stressors (Day 7 vs 8 and Day 7 vs 9) taken together), as illustrated below:

Measure Type	Training	Stressors	Overall
Control input	26%	56%	38%
System Performance	72%	37%	58%
Elapsed Time	2%	7%	4%

Control input (stick, pedal and throttle) measures represented 26 percent of the minimum measures during training and 56 percent of the minimum measures which describe performance changes due to task stressor changes.

Although it was important information that turbulence alone and interacting with aft C.G. caused a measure set change primarily in the control input measures, there was no rational way to justify the use of the resulting discriminant function for control of automated training under these conditions. Since turbulence alone and with aft C.G. were not measured early in training, the discriminant function could not be sensitive to the skill change *throughout training* for these conditions. Therefore, the recommended measures and weights which follow were restricted to light or heavy (aft C.G.) aircraft conditions.

RECOMMENDED MEASURES AND WEIGHTS

DISCRIM SELECT was altered to perform an analysis of the recommended measures for the purpose of stabilizing the beta weights. The results of the "ridge" reanalysis are shown alongside the non-biased results in tables 11 through 15 with the bias (k) value shown. Values from 0.1 to 0.5 were found to reduce the exaggerated weights as much as 70-80% without affecting the canonical R^2 or CHI-squared significantly, or altering the group means and standard deviations in discriminant space significantly (table 15).

The resulting weights from data biasing were considered the most stable model for use in the automated training system. The weights can be used directly to sum the measures into a single score for use by the adaptive logic (which requires a single score for performance assessment).

TABLE 11. RECOMMENDED MEASURES AND WEIGHTS FOR
MANEUVER 1, STRAIGHT AND LEVEL

SEGMENT	MEAS	Light A/C			Heavy A/C		
		K=0.0 ¹	K=0.4	COMM ²	K=0.0	K=0.4	COMM
Whole	ELRG	.997	1.186	.84	.985	.907	.74
Trial	ELF1	8.608	1.024	.25			
	AIRG	.374	.502	.76	-.109	-.011	.46
	PDRG	3.027	2.451	.35	7.940	5.671	.40
	PTRG	-.442	-.456	.80			
	PTSD				-1.440	-.719	.71
	ROAA				.271	.279	.52
	PSRM	.076	.142	.47	.265	.279	.46
	PSRG	-.023	-.047	.45	-.152	-.131	.49
	HAA	12.441	1.193	.64	8.259	2.131	.46
	HRG	-4.829	-.321	.71	3.512	2.342	.62
	HDAA	2.780	2.108	.82			
	ASAA	.012	-.0001	.48	.058	.109	.53
	ASRG	.049	.050	.63	.045	.040	.57
	R ²	.731	.688		.455	.412	
	X ²	336	298		170	149	

¹K is the bias added to the diagonal of the "within" matrix in the discriminant analysis to stabilize the weights. Where two values are shown, the recommended weights are below the highest k-value.

²Communality (the amount of variance of each measure extracted by all discriminant functions) shown is associated with the highest k-value.

TABLE 12. RECOMMENDED MEASURES AND WEIGHTS FOR
MANEUVER 2, CLIMBS AND DIVES

SEGMENT	MEAS	Light A/C			Heavy A/C		
		K=0.0	K=0.4	COMM	K=0.0	K=0.4	COMM
Whole	ELF1	10,640	1.120	.43			
Trial	ELF2	-1.753	.009	.74	2.528	1.069	.62
	ALRG	-.512	.078	.80	.167	.182	.67
	ALSD	5.403	1.871	.91	-.673	.269	.81
	PTSD	3.281	2.958	.92	2.041	1.301	.79
	HDAA	-1.955	-.453	.76	-.801	-.332	.49
	AIF2	3.527	.611	.50			
	ROAA	-.004	-.002	.47	.021	.030	.53
	PDF2	-.306	.276	.31			
	PSAA	.156	.170	.48	.095	.120	.41
	TURM	-.022	.017	.74	.451	.464	.78
	ASAA	.024	.040	.48	.117	.116	.51
R ²		.671	.637		.604	.599	
χ^2		285	259		261	257	

See footnote, table 11.

TABLE 13. RECOMMENDED MEASURES AND WEIGHTS FOR
MANEUVER 3, LEVEL TURNS

SEGMENT	MEAS	Light A/C			Heavy A/C		
		K=0	K=.5	COMM	K=0	K=.5	COMM
Whole	ELF2	-1.255	.075	.47			
	ALRG	-.449	.188	.73	-.165	-.126	.35
	ALSD	8.670	4.362	.82			
	PTSD	-.107	.141	.71	1.560	1.466	.68
	AIF2	-2.407	-.556	.41			
	ROAA	-.043	-.052	.06	.147	.230	.31
	PDF2	5.319	1.570	.28	14.465	5.578	.10
	ASAA	.172	.160	.67	.233	.237	.74
	HAA	-3.802	-2.359	.40	-4.202	-2.047	.53
R ²		.586	.528		.426	.365	
X ²		227	193		157	128	

See footnote, table 11.

TABLE 14. RECOMMENDED MEASURES AND WEIGHTS FOR
MANEUVER 4, CLIMBING AND DIVING TURNS

SEGMENT	MEAS	Light A/C			Heavy A/C		
		K=0	K=.3	COMM	K=0	K=.5	COMM
Initial	ELF1	-96.020	-.341	.20			
Climb	ALRG	.466	.413	.72	.081	.164	.45
or	HDAA	1.912	1.584	.76	4.666	4.251	.78
Dive	THRG	.031	.020	.45	-.033	-.021	.25
	ASAA	.158	.175	.13	.209	.237	.53
	ROAA	-.124	-.148	.03	.035	.045	.10
	PDF2				6.919	2.776	.14
R ²		.516	.495		.552	.516	
χ ²		189	179		227	206	
Climb	ELF2	-.183		.07			
or	ALRG	.470		.62	.101		.37
Dive	HDAF	.721		.36	.779		.18
Reversal	AIF2				-1.497		.24
	BERG	-.357		.18			
	TIME	.070		.59	.069		.68
	PDF1				-3.131		.15
R ²		.585			.342		
χ ²		161			93		
Final	ALRG	.472		.72	.085		.47
Climb	HDAA	.834		.79	1.611		.83
or	ASAA	.051		.60	.088		.75
Dive	ROAA	.075		.26	.065		.28
	PSAF	.019		.27			
R ²		.629			.274		
χ ²		257			91		

See footnote, table 11.

K values for the last two segments did not change weights materially.

TABLE 15. MEANS AND STANDARD DEVIATIONS OF DISTRIBUTIONS IN DISCRIMINANT SPACE¹.

MANEUVER	LIGHT A/C		HEAVY A/C	
	k=0	k=0.4	k=0	k=0.4
1. MEAN ²	1.467	1.49392	2.004	2.22700
S.D.	.775	.54307	.799	.74903
		<u>k=0.4</u>		<u>k=0.4</u>
2. MEAN	1.811	1.82923	2.550	2.55531
S.D.	.642	.58985	.760	.62989
		<u>k=0.5</u>		<u>k=0.5</u>
3. MEAN	1.481	2.08675	3.998	3.21209
S.D.	.642	.66571	1.331	.78041
		<u>k=0.3</u>		<u>k=0.5</u>
4-2. MEAN	.621	1.46996	3.487	3.23179
S.D.	.694	.70715	.976	.67713
		<u>k=0.0</u>		<u>k=0.0</u>
4-3. MEAN	1.666	1.66642	1.033	1.03348
S.D.	.642	.64270	.809	.80964
		<u>k=0.0</u>		<u>k=0.0</u>
4-4. MEAN	2.488	2.48889	2.192	2.19262
S.D.	.607	.60770	.850	.85043

¹See footnote, table 11.

²Group mean in discriminant space shown is for the criterion days. Group means for early training were negative and of equal magnitude since the two group discriminant analysis created a symmetrical coordinate system in discriminant space.

DISCUSSION

MEASURE SET COMPOSITION. The measure selection analysis data made two critical points which have an enormous impact on the design of performance measurement systems for automated flight training.

First, control input measures contained a significant amount of information about training and the effects of two task stressors. Typically, control input measures are not found in many training device measurement systems. Even advanced systems, such as the Automated Instrument Flight Maneuvers trainer, do not evaluate control inputs, primarily because without empirical data, such as those contained herein, it has been difficult to assess the implication of control input measures. The discriminant analysis removes some of these difficulties by not only selecting measures, but also assigning weights for the utilization of measures.

The second critical point has been seen in every measurement study conducted to date by the authors: Different measure sets are required when the task changes, even with the simple addition of light turbulence. Measure set composition changes alter both (a) the specific measures selected for each task, and (b) the weighting coefficients for these measures if the data are being summed into a single score.

Measures which are not useful for one condition, but which are "carried along" to cover a second condition, might degrade the power of the set to describe the first condition. Thus, one must be cautious in the application of measure sets to cover a variety of task situations. To guard against degrading the power of measurement, only empirical measurement studies offer an avenue to assure proper measure selection and compatibility at this time.

MEASURE SEGMENT START/STOP LOGIC. Existing computer programs were used to produce performance measures for the present study. In spite of their broad capacity to define when measurement segments start and stop, considerable testing was required to derive a set of logical conditions for starting maneuver 4, segment 4, the final climbing or diving turn.

The basic problem appeared to be that the existing logic tested for achieving several criterion conditions *simultaneously*. The logic did not permit the following kind of desirable expression: (a) Look for a 1000 foot altitude change. Then, after *a* has been found, stop looking for *a* and look instead for (b) altitude to return to within 1000 feet of the initial value. (c) When *b* becomes true, start measuring. If evaluated simultaneously, *a* and *b* would be mutually exclusive.

A new type of logical operator appears desirable in the start/stop logic, especially for maneuvering flight. Sequential .AND. (.SAND.) is proposed to join together any two logical or conditional expressions so that given that the first becomes true, testing of the first stops and testing of the second expression starts.

EQUIVALENT MEASURES ANALYSIS. The specification of initial candidate measures is a direct function of the skill of the analyst. Two kinds of measure specification errors have a high probability of occurrence. The most probable error appears to be overmeasurement. In the face of uncertainty caused by sparse evidence, the tendency is to adopt the philosophy, "If it moves, measure it." The second kind of error is to omit an important information form, such as control input.

These two kinds of errors represent a dilemma for the measurement analyst. If the candidate measure sets are terse, the risk of missing important information is high. Yet, if the candidate sets are abundant, the risk and cost of overmeasurement can be so enormous that data collection becomes impractical. Even if data collection is possible, the multivariate procedures for measure selection require seven to nine times as many data points as input variables to work properly; data collection requirements are a direct function of the number of measures initially specified.

The use of correlation analysis to reduce redundant forms of information appears to be a useful tool to ease the dilemma. It serves as a first step check on the analyst. Also, it permits the analyst a little latitude to experiment with candidate measures in selected areas of uncertainty. However, heavy dependence on the equivalent forms analysis to eliminate redundant measures should be avoided.

MINIMUM STATISTICAL SAMPLE. The discriminant analysis technique appears to have been effective, but considering the amount of data which were collected, it may be wondered if a smaller statistical sample could have sufficed.

A relatively small number of participants were used (12). The adequacy of this sample depends on the population to which one wishes to extrapolate. On the other hand, a large amount of data was collected from these participants over a quantity of experimental trials (5184 TOTAL). The technique used in this study would be more easily applied in the future if the amount of data collection could be reduced. Consequently it is appropriate to ask if sufficient statistical power would be maintained with fewer observations.

(One may attempt to control two types of errors in the design of an experiment: (a) the error of asserting that a result is "real" when in fact it occurred by chance, and (b) the error of asserting that a result occurred by chance when in fact it was "real." The probability of the first type of error is controlled by the use of a table of statistics corresponding to the desired probability of chance occurrence. The probability of the second type of error is controlled by the use of a sufficient number of observations.

The necessary sample to achieve the required double conditions can be determined for the F-test and one-way comparisons by: (a) specifying the minimum practically important differences one wishes to detect, (b) determining the experimental error which will be encountered, and (c) based on α and b , reading the needed sample from published tables (cf., Winer, 1962, pp. 657-658; Scheffe, 1959, pp. 438-455). Now that data are on hand, it is possible to conduct such an examination with minor computer program modification and re-analysis.

(It is possible, also, to modify the computer programs and repeat the analyses with the amount of data successively reduced to empirically find the minimum allowable sample. Such re-analysis should permit future applications of the techniques of this study with increased efficiency.

"RIDGE" WEIGHT STABILIZING. The values of k selected to stabilize the weighting coefficients were higher than those typically used in the literature, and may generate controversy among mathematical statistitions. It was noted, however, that the discriminating power of the measure set was not significantly changed by the high k values. A partial validation of the technique resulted when the recommended measures and weights generalized to a new subject sample in Phase III.

(AVERAGE WEIGHTED SCORES. The group centroids in discriminant space cannot be used directly to establish the expected "average" when raw data are weighted and summed. This is because the discriminant analysis transforms the data so that they are symmetrical in discriminant space. The data base must be recalculated using recommended measures and weights to establish the actual means and standard deviations on the untransformed discriminant function (as can be seen in the next section on measurement implementation).

SECTION IV

MEASUREMENT IMPLEMENTATION

Phase II was a computer software re-design and implementation effort which required some measurement data re-analysis and the development of scaling methods to relate new measurements to the existing adaptive logic. An engineering system test was conducted to insure that the system did work during training conditions.

APPARATUS

The experimental flight simulator was the TRADEC, located at the Naval Training Equipment Center. TRADEC was converted into an automated instrument flight trainer as described in Section II.

TRAINING COURSE

The original IFM training course consisted of 65 different exercises which contained 18 straight and level exercises, 20 climbs and dives, 15 level turns and 12 climbing and diving turns which were ordered in increasing levels of "difficulty." The inherent complexity of the maneuvers was one factor of difficulty. Other difficulty factors were changes in aircraft weight (center of gravity) and drag, atmospheric turbulence and the speed at which the aircraft was flown.

An analysis of the original IFM training course suggested that some of the maneuver task combinations were not necessary for measurement evaluation purposes. The training course was shortened to 44 different exercises which contained 8 straight and level runs, 12 climbs and dives, 12 level turns and 12 climbing and diving turns listed in table 16. The modified course contained the fundamental elements of the original task maneuvers and two combinations (each) of aircraft center-of-gravity and airspeed.

Aircraft weight and center-of-gravity shift from a more stable longitudinal axis control task to a less stable one was obtained by manipulating fuel and external stores. The following conditions are shown in table 16:

C.G. Level 1	2400 pounds internal fuel.
C.G. Level 2	2 Sidewinder missiles (stations 2, 8), full internal fuel and full center-line tank.

Two airspeeds were used as shown in table 16. The slower speed, 280 knots, was more difficult to fly than the higher speed because of aircraft stability differences.

TABLE 16. MODIFIED SYLLABUS

MAN	SEQ #	IFM #	BANK ANGLE	IAS Kts	TURN DEGREE	CLIMB FEET/MIN	CG
STR & LVL	01	01	0°	350	0	0	1
	02	01	0°	350	0	0	1
	03	01	0°	350	0	0	2
	04	01	0°	350	0	0	2
	05	04	0°	280	0	0	1
	06	04	0°	280	0	0	1
	07	04	0°	280	0	0	2
	08	04	0°	280	0	0	2
CLB & DIV	21	32	0°	350	0	-1000	1
	22	33	0°	350	0	1000	1
	23	32	0°	350	0	-1000	1
	24	33	0°	350	0	1000	2
	25	32	0°	350	0	-1000	2
	26	33	0°	350	0	1000	2
	27	36	0°	350	0	-1000/+1000	1
	28	37	0°	350	0	1000/-1000	1
	29	36	0°	350	0	-1000/+1000	1
	30	37	0°	350	0	1000/-1000	2
	31	36	0°	350	0	-1000/+1000	2
	32	37	0°	350	0	1000/-1000	2
LVL TRN	51	55	30°	350	45	0	1
	52	56	30°	350	-45	0	1
	53	55	30°	350	45	0	1
	54	56	30°	350	-45	0	2
	55	55	30°	350	45	0	2
	56	56	30°	350	-45	0	2
	57	58	30°	350	90/-90	0	1
	58	66	30°	350	-90/+90	0	1
	59	58	30°	350	90/-90	0	1
	60	66	30°	350	-90/+90	0	2
	61	58	30°	350	90/-90	0	2
	62	66	30°	350	-90/+90	0	2
CLB & DIV TRN	71	71	30°	280	90	-1000	1
	72	72	30°	280	-90	1000	1
	73	73	30°	280	-90	-1000	1
	74	74	30°	280	90	1000	2
	75	73	30°	280	-90	-1000	2
	76	74	30°	280	90	1000	2
	77	79	30°	280	90/-90	1000/-1000	1
	78	80	30°	280	-90/+90	-1000/+1000	1
	79	79	30°	280	90/-90	1000/-1000	1
	80	80	30°	280	-90/+90	-1000/+1000	2
	81	79	30°	280	90/-90	1000/-1000	2
	82	80	30°	280	-90/+90	-1000/+1000	2

At the beginning of each exercise, the computer program would set the aircraft at the initial run conditions, brief the trainee and turn simulator control over to the trainee when the trainee acknowledged instructions. All runs that did not involve reversals in turn directions or reversals in vertical path were nominally one minute in duration. All runs that required reversals were nominally two minutes in length, although it might take the trainee longer than the nominal time to perform them. Limits were placed in the program to stop exercises if reasonable times to perform were exceeded. Crash or completely out of control conditions also stopped the exercise.

ORIGINAL IFM PERFORMANCE SCORING

IFM contained a performance measurement module and an adaptive logic that permitted the student to sequence through the training course according to his/her measured performance. Since the adaptive logic was based on measurement assumptions, it is important to the present effort to review the rationale behind the original scoring algorithm.

The performance measurement parameters were developed from NATOPS standards for instrument flight in accordance with the performance band limits shown in table 17. It was assumed that the NATOPS middle bandwidth represented a 95% probability (+2 standard deviations), and that this level of performance denotes acceptable performance for an experienced naval aviator (Charles, et al, 1972). Thus, 95% of all performance by experienced aviators would fall within the middle bandwidth. It was further assumed that any error data about the nominal values were normally distributed, and that the inner bandwidth represented one standard deviation (about 68% of performance) and the outer bandwidth represented four standard deviations (100% of performance).

Error from the desired values of three parameters were obtained during execution of each maneuver, as shown in table 18. Parameters were sampled twice per second, subtracted from the desired value, multiplied by a normalizing constant (see table 19.), summed into a root-mean-square error score across all three parameters for the entire run length, then divided by the proportion of the run completed as shown in table 20.

The resulting error score was positive in value and increased with poor performance. A total score of 75 would indicate, for example, that all three parameters were held at the inner band limits for the entire run (eg heading at 5°, altitude 100' and airspeed at 5 kts). According to the scoring rationale, this would represent one standard deviation performance. Similarly, a total score of 150 would be representative of a run in which all three parameters were held at middle performance limit, or two standard deviation performance.

TABLE 17. ORIGINAL IFM PERFORMANCE BAND LIMITS

PARAMETER	INNER	MIDDLE	OUTER
Heading	$\pm 5^{\circ}$	$\pm 10^{\circ}$	$\pm 20^{\circ}$
Altitude	$\pm 100'$	$\pm 200'$	$\pm 400'$
Airspeed	± 5 Kts	± 10 Kts	± 20 Kts
Vertical Velocity	$\pm 250'/\text{Min}$	$\pm 500'/\text{Min}$	$\pm 1000'/\text{Min}$
Turn Rate	$\pm 0.5^{\circ}/\text{Sec}$	$\pm 1.0^{\circ}/\text{Sec}$	$\pm 2.0^{\circ}/\text{Sec}$
Bank Angle	$\pm 2.5^{\circ}$	$\pm 5^{\circ}$	$\pm 10^{\circ}$

TABLE 18. ORIGINAL IFM PARAMETERS SCORED

MANEUVER	PARAMETER					
	HEAD- ING	BANK ANGLE	TURN RATE	ALTI- TUDE	RATE OF CLIMB	IAS
Straight & Level	X			X		X
Climbs & Dives	X				X	X
Level Turns						
Fixed Angle		X		X		X
Fixed Rate			X	X		X
Climbing and Living Turns			X		X	X

TABLE 19. ORIGINAL IFM WEIGHTING COEFFICIENTS

PARAMETER	COEFFICIENT (K)
Heading	5.00
Altitude	0.25
Airspeed	5.00
Vertical Velocity	0.10
Turn Rate	50.00
Bank Angle	10.00

TABLE 20. ORIGINAL IFM SCORING ALGORITHM

$$S_e = K (P_c - P_a)$$

where:

S_e = parameter error score

P_c = desired value of parameter, P

P_a = actual value of parameter, P

K = parameter normalizing constant

and:

$$S_t = \frac{\sum_{i=1}^3 \sqrt{\frac{\sum_{j=1}^N (S_{e_j})^2}{N}}}{P_r}$$

where:

S_t = total score for run

S_e = error score for each of three parameters sampled

N = number of samples

P_r = proportion of run time completed in seconds, $\frac{I_t}{\text{actual time}}$

I_t = ideal time, the time required to complete a perfect maneuver

ORIGINAL IFM ADAPTIVE LOGIC

Based on the measurement algorithm, an adaptive logic was developed to permit the student to advance through the training course. The logic is shown in table 21. At the end of a given run, a student will either advance one, two or three numbered exercises in the sequence, stay the same, or go back one, two or three exercises as a function of his current performance and whether or not he had advanced or moved backwards in the syllabus on his previous run.

TABLE 21. ORIGINAL IFM ADAPTIVE LOGIC

Previous Run Sequence Number Increment Status	S_t >200	$200 > S_t$ >150	$150 > S_t$ >100	$100 > S_t$ >50	$50 > S_t$
-(Decrement)	-3	-2	0	0	+1
0 (No Change)	-2	-1	+1	+1	+2
+(Incremented)	-1	0	+1	+2	+3

MAPPING NEW MEASURES INTO EXISTING ADAPTIVE LOGIC

For subsequent measurement evaluation purposes it was necessary to have three scoring systems, (1) the original IFM scoring system, (2) a system based on DISCRIM recommended measures and weights and (3) a system based on observed, normative IFM scores. A method to scale the second and third measurement systems into the adaptive logic was required.

SCALING METHOD. In the original IFM adaptive logic (table 21), the decision to branch was made on the basis of the assumed distribution of the total IFM score, S_t , where:

$S_t = 75$ was assumed to be 1-sigma performance, and

$S_t = 150$ was assumed to be 2-sigma performance for the experienced naval aviator.

Therefore, branching decisions can be expressed as a function of score standard deviations (z-scores), as follows:

$$S_t = 50 = .667\sigma,$$

$$S_t = 100 = 1.333\sigma,$$

$$S_t = 150 = 2.000\sigma \text{ and}$$

$$S_t = 200 = 2.667\sigma.$$

By computing z-scores, the second and third measurement system scores can be scaled into the existing adaptive logic without changing the rationale upon which the adaptive logic was designed.

DISCRIM MEASUREMENT SCALING. The Phase I data were recomputed using the recommended measures and weights (tables 11 - 15). On each trial, each recommended measure was multiplied by its respective weight. A single score for each trial was computed by summing the weighted measures. This new metric for each trial was called the *total weighted score* (S_{tw}). The average total weighted scores are shown in table 22 along with their standard deviations for each maneuver and segment.

For purposes of establishing a z-score, criterion data were drawn from Day 7 for light aircraft and from Day 6 for a heavy aircraft. Thus for every trial of a given maneuver (and for each segment within a maneuver), a score would be computed for evaluation by the adaptive logic as follows:

$$S_z = \left| \frac{S_{tw} - S_{twcm}}{S_{twcs}} \right|$$

where,

S_z = the total score expressed as the absolute value of standard deviations of criterion performance,

S_{tw} = the total weighted score for each segment,

S_{twcm} = the S_{tw} mean performance on the criterion day,

S_{twcs} = the S_{tw} standard deviation on the criterion day.

Where maneuvers contained more than one segment, the S_z value passed to the adaptive logic would be the average of all S_z values. If any segment failed to start or stop, S_z would be set to 2.700-sigma for that segment.

During system engineering tests it was discovered that negatively weighted measures could cause misclassification of exceptionally poor performance (such as turning the wrong way). In each case, the poor performance was found to be way outside of the measurement space of the Phase I data. The maximum values for all negatively weighted measures observed in the Phase I data base are shown in table 23.

To guard against the possibility of misclassification by the discriminant scoring model, all negatively weighted measures were first tested against the limits in table 23. If on any trial a negatively weighted measure was greater than its limit, S_{tw} was not computed, and a constant S_z of 2.700 was returned to the adaptive logic.

TABLE 22. AVERAGE TOTAL WEIGHTED SCORES FOR USE IN
NEW MEASUREMENT SYSTEM

MAN		LIGHT A/C		HEAVY A/C	
		DAY 1	DAY 7	DAY 2	DAY 6
1	MEAN	3.16813	1.49352	3.54622	2.22700
	S.D.	.54307	.54307	.74903	.74903
	N	132	132	144	144
2	MEAN	3.43938	1.82923	4.10383	2.55531
	S.D.	.58985	.58985	.62989	.62989
	N	132	132	144	144
3.	MEAN	3.57394	2.08675	4.45655	3.21209
	S.D.	.66571	.66571	.78041	.78041
	N	132	132	144	144
4-2	MEAN	2.87751	1.46996	4.69867	3.23179
	S.D.	.70775	.70775	.67713	.67713
	N	133	133	144	144
4-3	MEAN	3.19163	1.66642	2.10973	1.03348
	S.D.	.64270	.64270	.80964	.80964
	N	94	94	114	114
4-4	MEAN	4.07226	2.48889	3.23810	2.19262
	S.D.	.60770	.60770	.85043	.85043
	N	132	132	144	144

TABLE 23. UPPER BOUNDS OF NEGATIVELY WEIGHTED MEASURES

MANEUVER SEGMENT	MEAS	LIGHT A/C	MEAS	HEAVY A/C
1	PTRG	6.780	AIRG	2.010
	PSRG	7.700	PTSD	1.607
	HRG	355.000	PSRG	7.136
	ASAA	13.578		
2	HDAA	623.000	HDAA	780.000
	ROAA	5.094		
3	AIF2	0.656	ALRG	7.259
	ROAA	13.365	HAA	176.000
	HAA	281.000		
4-2	ELF1	0.017	THRG	5.435
	ROAA	31.498		
4-3	ELF2	2.710	AIF2	1.337
	BERG	3.428	PDF1	1.136
4-4	----	----	----	----

NORMATIVE IFM MEASUREMENT SCALING (NEW IFM). Original IFM scores (S_t) were collected during Phase I. It was observed that the data from the subject sample did not agree with the assumed performance norms (ie average performance was assumed to be $S_t=75$). Table 24 suggests that the original IFM adaptive logic was too lenient for straight and level flight and too demanding for climbing and diving turns based on Day 6 and Day 7 data. Since the original IFM measurement represented an analytically specified, criterion referenced measurement system based on performance norms of IFM measurement for subsequent evaluation.

The IFM scores from Phase I had the characteristics of a Poisson distribution; the mean represented 1-sigma performance. Day 6 and Day 7 means for each maneuver of C.G. condition were multiplied by 0.667, 1.333, 2.000, and 2.667 to determine the adaptive logic decision values shown in table 25. From a programming viewpoint, it was easier to replace the decision values than to compute z-scores for NEW IFM scoring. The result was equivalent. Thus all three measurement systems were scaled into the adaptive logic in an equivalent manner.

TABLE 24. AVERAGE IFM SCORES FROM PHASE I

MANEUVER	LIGHT A/C		HEAVY A/C	
	DAY 1	DAY 7	DAY 2	DAY 6
STRAIGHT & LEVEL	69 ¹	34	55	34
CLIMBS & DIVES	125	50	144	57
LEVEL TURNS	146	65	121	68
CLIMBING & DIVING TURNS	221	94	206	120

¹N = 144

TABLE 25. ADAPTIVE LOGIC FOR ALL SCORING SYSTEMS

SCORING SYSTEM	LOGIC DECISION VALUES AND RESULTING RUN INCREMENT OR DECREMENT (BELOW)				
I. ORIGINAL IFM	$S_t > 200$	$>S_t > 150$	$>S_t > 100$	$>S_t > 50$	$>S_t$
II. DISCRIM	$S_z > 2.667$	$>S_z > 2.000$	$>S_z > 1.333$	$>S_z > 0.667$	$>S_z$
III. NEW IFM	$S_t > (1)$	$>S_t >$	$>S_t >$	$>S_t >$	$>S_t$
Fore C.G.	↓	↓	↓	↓	
Man. 1	91	68	45	22	
Man. 2	133	100	67	33	
Man. 3	173	130	87	43	
Man. 4	251	188	125	62	
Aft C.G.					
Man. 1	91	68	45	22	
Man. 2	152	114	76	38	
Man. 3	181	136	91	45	
Man. 4	320	240	160	80	
<hr/>					
Previous Run Sequence Status					
- (Decrement)	-3	-2	0	0	+1
0 (No Change)	-2	-1	+1	+1	+2
+ (Increment)	-1	0	+1	+2	+3

¹The criterion of S_t shown below for each maneuver (man.) and C.G. condition to be inserted here.

MEASUREMENT IMPLEMENTATION

The IFM system computer programs were modified to incorporate the new measurement systems and to permit subsequent measurement evaluations. Considerable programming was required. The modifications to specific program modules are outlined in Appendix D. A summary of those modifications follows:

1. The instruction syllabus was shortened as shown in table 16. Basically, an intermediate level of c.g. and all turbulence conditions were removed. Also, some unnecessary combinations of climbing and diving turns were eliminated.
2. Real-time plotting of IFM measure time histories on the IIDOM was removed from the program to decrease operating complexity and increase storage space.
3. Maneuver segmentation for measurement purposes was added. The segmentation algorithms included the logical operators and conditional test functions described in previous measurement work.
4. The capability of sampling each parameter at a unique rate was added.
5. Each measure was defined as a parameter, desired value and transform, per previous work.
6. The S_{tw} measurement algorithm and limit tests were added.
7. The S_z measurement algorithm was added.
8. The NEW IFM measurement algorithm was added.
9. The program was modified to operate either according to the old IFM, DISCRIM or NEW IFM measurement systems by selecting sense switch options.
10. The performance summary line printer output was modified to include DISCRIM measures in their raw form, weighted measures, the sum of weighted measures (S_{tw}), S_z , the criterion S_{tw} (where multiple segments exist).
11. A tape writing module was created to output all subject and performance data on magnetic tape at the end of each trial.

SYSTEM TEST PROCEDURES

A system check-out was conducted to insure that the program was working properly, that measures were being properly sampled, transformed, weighted and acted-upon properly, that the maneuver segmentation rules worked, that the line printer output was correct, and that sufficient foreground processing time resulted. This test was not intended to be any kind of a system evaluation.

The tests were conducted informally by checking-out each module change as applicable, and by flying the system with each measurement system controlling training. Two test trainees were used; they were low-time private pilots who had only light aircraft experience. Testing with the second trainee revealed the potential misclassification problem with the initial DISCRIM measurement system (previously discussed) and brought about solution to that problem.

SECTION V

MEASUREMENT SYSTEM EVALUATION

The purpose of Phase III was to conduct a pilot study to evaluate measurement development progress to date. The three measurement techniques which resulted from Phase I and II were evaluated by empirical comparison of the time-to-train (to criterion) three groups of six novice pilots each using the original IFM (Group I), discriminant (Group II) or normative IFM (Group III) scoring subsystems in IFM.

METHOD

APPARATUS. The TRADEC and automated Instrumented Flight Maneuvers (IFM) program was made to operate with three scoring subsystem described in Section IV.

TRAINEES. Fifteen, 17 to 40 year old, light aircraft, civilian pilots were used. An attempt was made to restrict the pilot sample to high-time Student Pilots or low-time Private Pilots who had between one and five hours of instrument time. It was expected that this sample would approximate the population that might benefit from IFM automated training.

MATCHING GROUPS. In addition to the above criteria, pilots were divided into three equivalent groups, matched on two variables, recency and first run IFM scores. Recency was calculated as follows: The total of hours flown in the last 10 days plus hours flown in the last two months was divided by 10. The second variable was the first IFM trial score after initial practice.

It was not possible to test all pilots for group assignment at one time because of the uncertainty of volunteer pilot schedules over the 10 weeks required to collect data. Matching was done when pilots arrived for their first session by assignment to keep running means of first scores and recency as equivalent as possible. Of course, the degrees-of-freedom to accurately match reduced as the experiment progressed.

PROCEDURES. At the first session the test conductor briefed the trainee on the purpose of the study, use of the data, the TRADEC flight instruments and controls, the differences between high performance aircraft and light aircraft and the study procedures. Each pilot was given between one and three practice trials to demonstrate ability to control the simulator.

The pilot was then selected for one of the three scoring systems using the matching method, and given a sequential subject number within group (ie Subject 3, Group 2). All trainee data and performance records were indexed only by subject and group number. There was no way to link the data records to a specific person without knowing his/her subject and group number; the

index between subject/group number and individuals was destroyed at the end of the study.

After group assignment, the trainee was placed under full control of the automated training system. Pilots were permitted to fly 45 to 50 minutes under control of their assigned scoring system. On successive days, the training syllabus was started with the last exercise flown on the previous day. Training continued until the last exercise was flown and a passing score was achieved.

MEASUREMENT. Since automated IFM trained to criterion (as expressed by the measurement and adaptive logic described in Section IV), the dependent variable was the number of trials required or the time-to-train, used interchangeably, to complete the course. Both IFM and S_z scores (see Section IV) were available to assess performance quality as well.

RESULTS

MATCHING GROUPS. There were no statistically significant differences between groups for trainee data shown in table 26. Inspection of the distributions and trends, however, suggested that recency and total flight time favored Group I. First trial IFM scores favored Group III. Age favored Group II. S_z , Discrim scoring, was not sensitive for matching at this initial stage of training; many scores of 2.700 indicated that the model measurement space was exceeded during initial matching runs.

RAW RESULTS. There were no significant differences between groups on the last trial IFM or S_z scores; groups were trained to equivalent performance levels (table 27). The number of trials to achieve this performance was significantly different for Group II, representing a 72% reduction in the time-to-train over Group I. Group III was not significantly different from Group I or II. It was suspected, however, that these results may have been biased by imperfect group matching.

VARIABLES AFFECTING GROUP COMPOSITION. Correlations were calculated between the variables shown in table 28. Group membership was set to 0 for Group I, to 2 for Group II, and to 1 for Group III (in order of performance) for correlation analysis purposes. Group membership correlated with number of trials with an $r = -.47$, accounting for only 22% of the variance in the data. The partial correlation between groups and trials, holding first score constant was $r_{gt.f} = -.51$. The partial correlation between groups and trials holding age constant was $r_{gt.a} = -.39$. Age and first score were biasing the data.

A stepwise multiple regression (Heal, 1971) was performed with variables one through six available as predictors; variable seven (trials) was the criterion. The stepwise process permitted only significant predictors to enter the model, based on preset F-ratio criteria. The F-level required to enter or be rejected

TABLE 26. TRAINEE DATA

	GROUP	RECENCY	TOTAL ¹ TIME	AGE	INST ² TIME	FIRST TRIAL IFM	SCORES S _z
I.	MEAN	2.47	327.3	28.20	4.9	150.35	2.50
	S.D.	1.24	627.7	6.38	8.5	91.61	.44
II.	MEAN	.69	89.0	23.80	2.4	154.45	2.70
	S.D.	.74	44.8	9.31	2.1	82.84	.00
III.	MEAN	.96	97.5	28.00	5.9	112.71	2.38
	S.D.	.72	74.1	1.41	7.9	42.04	.71

¹Total flight time in hours.²Total instrument time in hours.

TABLE 27. RAW RESULTS

	GROUP	FIRST TRIAL	LAST TRIAL IFM	S _z	RAW TRIALS	PERCENT IMPROVEMENT
I.	MEAN	150.35	117.90	1.10	98.20	
	S.D.	91.61	58.45	.74	48.49	
II.	MEAN	154.45	95.92	1.07	56.80 ¹	72%
	S.D.	82.84	27.08	.22	28.22	
III.	MEAN	112.71	128.76	1.35	62.20 ²	57%
	S.D.	42.04	68.89	.41	21.20	

¹Significant, Mann-Whitney U=3, p=.028.²Not Significant.

TABLE 28. SIMPLE CORRELATIONS BETWEEN VARIABLES¹

VARIABLE	1	2	3	4	5	6
1. RECENCY	1.00					
2. TOTAL TIME	.56	1.00				
3. AGE	.01	-.10	1.00			
4. INST TIME	.35	.62	-.04	1.00		
5. FIRST SCORE	-.26	-.39	-.21	-.27	1.00	
6. GROUPS	-.46	-.28	-.29	-.16	.02	1.00
7. TRIALS	-.04	-.21	.68	-.26	.34	-.47

¹ $r = .514$ sig., $p = .05$, two tailed, $r = .414$ for one tailed.

TABLE 29. MULTIPLE REGRESSION RESULTS

PREDICTORS	β	b	STD ERROR	F
AGE	.7888	4.570	.082	11.43 ¹
FIRST SCORE	.5011	.258	.924	9.87 ¹
MULTIPLE R = .8415, $R^2 = .7082$				
CRITERION = NO. TRIALS				
PREDICTION EQUATION ²				
TRIALS = 4.570 AGE + 0.258 FIRST SCORE - 85.420				
STANDARD ERROR = 21.755				

¹Significant, $p < .001$, 2/12 df.

²Describes this data base only.

from the regression analysis was set to $F=3.59$, which would permit up to three predictors at 3/11 df.

Only two predictors entered the multiple regression, age and first score as shown in table 29. Age and first score taken together and weighted could predict the number of trials with a standard error of 21.76 (trials), without regard to group membership, and accounted for 70% of the variance in the data. With this result, the effects of age and first score could be partitioned, thereby statistically equating the groups on these significant variables.

EVALUATION RESULTS. Age and first trial effects were removed from the data by subtracting the number of predicted trials from the raw trials and forming a difference score (DIFF in table 30). The difference scores placed Group I performance 15 trials above the grand mean, Group II six trials below the grand mean and Group III nine trials below the grand mean. Both Groups II and III were significantly different from Group I.

The difference scores were added to the grand mean of trials to form an adjusted number of trials (ADJ TRIALS in table 30). With the effects of age and first trial scores thus removed, Group II produced a 34% reduction, and Group III produced a 40% reduction in the time-to-train over Group I.

On a maneuver by maneuver basis, Discrim scoring held trainees in straight and level flight longer than either IFM scoring systems (table 31). Discrim scoring permitted trainees to pass through climbs and dives and level turns faster than either IFM scoring system, and through climbing and diving turns faster than Old IFM scoring. Note that these data were based on raw (unadjusted) trials.

The performances of three typical pilots who were close to their group means are presented in Appendix E. These graphs plot the progress of each trainee through the syllabus by trial. They show that Discrim scoring tended to hold the trainee in the first exercise of straight and level flight much longer than either of the two IFM scoring systems. Also, both IFM scoring systems produced noticeably more instabilities (oscillations up and down the exercise list) than Discrim scoring.

Three subjects trained on Old IFM scoring volunteered comment that the scoring and adaptive logic seemed arbitrary a few times when their perceived performance did not agree with the automated judgments. No such comment was volunteered for the other two scoring systems.

All subjects had problems with the Cognitronics corrective messages during early training. The Cognitronics issued corrections when altitude, heading, airspeed, rate of descent or bank angle were out of tolerance. When multiple performance errors occurred, the corrective messages would "stack-up" in a queue,

TABLE 30. MEASUREMENT EVALUATION RESULTS

GROUP	FIRST TRIAL	AGE	RAW TRIALS	PREDICT TRIALS	DIFF	ADJ TRIALS	PERCENT IMPROV
I. MEAN	150.35	28.20	98.20	82.29	15.91	88.31 ³	
S.D.	91.61	6.38	48.49	37.81	12.71		
II. MEAN	154.45	23.80	56.80	63.23	-6.43 ¹	65.97	34%
S.D.	82.84	9.31	28.22	39.36	20.32		
III. MEAN	112.71	28.00	62.20	71.66	-9.47 ²	62.93	40%
S.D.	42.04	1.41	21.20	15.42	19.11		
GRAND MEAN			72.40	72.40	0.00	72.40	

¹GROUP II vs I, Mann-Whitney U=4, sig, p=.048.

²GROUP III vs I, Mann-Whitney U=0, sig, p<.001.

³Adjusted trials = DIFFERENCE + TRIAL GRAND MEAN.

TABLE 31. NUMBER OF RAW TRIALS TO COMPLETE EACH MANEUVER

GROUP	STRAIGHT & LEVEL	CLIMBS & DIVES	LEVEL TURNS	CLIMBING AND DIVING TURNS
I.	15 ⁽¹⁾	30	32	21
II.	22	16	9	9
III.	17	22	16	7

⁽¹⁾ Average No. trials, N=5 per group.

awaiting delivery of previous messages. Often a coaching message would occur after corrective action was taken, causing the trainee to overcorrect. Later in training the error rates were down and the trainees learned to ignore the messages.

DATA COLLECTION NOTES. Twenty-nine pilots volunteered for the study in response to notices given to three general aviation fixed base operators in the Orlando, Florida area. Eight volunteers were ruled-out because of very high total flight or instrument hours. Two potential trainees were excused after matching because they could not be assigned a Group without significantly unbalancing recency or first scores. One trainee started but did not finish due to continued conflict with his work schedule. When data collection was finished, there were 6 subjects in each group (N=18).

Three trainees over 40 years old were omitted (the oldest from each group) during preliminary analyses because (1) the age effect was more pronounced than anticipated, (2) they were outside the expected age range of potential automated IFM system users, (3) they were outside the age range of the Phase I data from which both Discrim and Normative IFM measurement "models" were derived, (4) their outlying performance introduced an unprecedented amount of variance in the data, and (5) in one case, the trainee did not appear to be very adaptable to automated training techniques as configured in IFM.

Data collection required 10 weeks, scheduling an average four hours of system time each day for an average of five days a week (M,T,T,F,S). About one hour a day (or one day a week) was lost due to trainee no-show, trainee late or system malfunction (in order of decreasing occurrence).

DISCUSSION

The results offered encouraging evidence that empirical methods can improve upon analytically derived measurement and cause a substantial increase in the efficiency of training. Flight simulators are scheduled heavily in the field. Present and future systems can be expected to be burdened with even higher utilization due to more training required by more complex systems, tasks and pressures to conserve fuel. A 40% increase in training efficiency would have a substantial impact in field training.

AGE. Subject age was a more powerful influencer of complex psychomotor training performance than the measurement systems, where the range of age in the sample was between 17 and 40 years. Although we did not need to perform a study to learn that, we had to be certain that age (and other variables) were not biasing the data in favor of one measurement system over another. The use of the prediction equation removed the bias, and was conservative because there was correlation (table 28) between age and groups (ie some of the group effect was removed by the procedure).

The magnitude of the age effect suggested that it should be included in any future application of the discriminant analysis measure selection technique. It should become one of the candidate measures along with other student history variables as well. Future studies of this type should match groups on age and first score.

VALUE OF NORMS. The time-to-train improvement for the normative IFM measurement group suggested the efficiencies that can be obtained by simply collecting empirical data and adjusting analytically derived measurement according to performance norms. This has implication for retro-fit or situations where the discriminant (or other multivariate) techniques are not feasible.

DISCRIM PARTIALLY VALIDATED. The discriminant model developed in Phase I generalized to a new sample of pilots who had 58% more total flying hours, 54% less instrument time and who were 16% older. It also trained as well as the "criterion referenced" normative IFM measures. This suggests some validity in the model as a whole, which included (1) removal of the components of variance to create "independent" samples, (2) the use of the multiple discriminant model for measure selection, and (3) the "ridge" method to stabilize the weights.

TOWARD MORE COMPREHENSIVE MEASUREMENT. The discriminant model did not perform any better overall than the normative IFM model. Discrim, however, has advantages that may lead to an improvement in efficiency beyond normative criterion referenced models. The principle advantage is that DISCRIM SELECT can accept non-system performance measures and properly weight and evaluate them in a set that contains also system performance measures.

For example, if pilot age had been included in the Phase I candidate measure set, it probably would have emerged as a recommended measure (based on Phase III results). If it had, the evaluation results would probably have been closer to the raw results (table 27) than the adjusted results because one of the measure groups would have been sensitive to the age effect, and would have absorbed some of the age effect variance. There are undoubtedly several student history variables that are just as important to performance assessment as the system performance measures.

PILOTING TECHNIQUE. The discriminant model essentially described a trained person in multidimensional space, which included control input measures as well as outer-loop (ie heading, altitude and airspeed) measures. It is possible that Discrim scoring was sensitive to pilot control technique as well as overall system performance.

REINFORCEMENT. Discrim scoring was not sensitive to performance differences during matching and held trainees in the very first straight and level exercise for a long time. Decisions made on Discrim scoring required that the trainee start performing like

a trained person in all dimensions (including control input) before it would permit any advancement. Pilots so trained did not receive any positive reinforcement (advancement) until they developed sufficient technique and performance. Once that happened, they progressed rapidly through the syllabus.

In contrast, both IFM scoring systems were less demanding; only outer-loop measures had to be within bounds. IFM scoring systems may have permitted advancement prior to the development of good basic piloting technique. Trainees may have been incorrectly rewarded by advancement (note comments by subjects that IFM seemed arbitrary at times) and were still trying to discover proper technique while encountering new tasks. This could have caused the instabilities that were seen in both IFM scoring systems.

SINGLE SCORE MEASUREMENT. We are not convinced that adaptive logics which require movement through a syllabus based on a single score produce the most efficient training. Performance is multidimensional, and measurement can be made to diagnose at least major problems. For example, if a student during a climbing turn has problems controlling the turn, that problem is easily measured. Diagnosis of the problem and subsequent action by the adaptive logic might produce more efficiencies and perhaps better training.

MEASUREMENT RELATIVE TO STUDENT EXPERIENCE. Early in training a student may not need to perform within 2-sigma of end of course criteria. If a student is within the performance range of other students with his experience (and those norms converge on end of course criteria), then the student is performing as expected, and should be permitted to advance. Adaptive logics can be designed to make judgments based on such norms. When the system is first installed it can start operation with assumed norms that can be programmed to adjust after sufficient data are accrued.

COMMENTS ON AUTOMATED TRAINING SYSTEM DESIGN. Although the purpose of our work was to develop and evaluate measurement, several comments on the design of automated training systems can be made on the basis of the training problems that were observed. These comments might be helpful to designers of next generation systems, and are contained in Appendix F.

SECTION VI

CONCLUSIONS

The purpose of the program was to develop improved measure selection techniques, implement the results of those techniques in an automated flight training system and evaluate resulting measurement. All program objectives were achieved. The conclusions of each phase are presented in the following:

MEASURE SELECTION

Noting that task analytic methods often produce an abundance of measurement, empirical techniques were explored to reduce analytically derived measurement to a smaller, more manageable set that would be sensitive to the change in performance during training. The major conclusions of the measure selection method development work were:

1. It is necessary to perform a good analysis of each maneuver to specify *candidate* measurement based on operational requirements and the research literature.
2. Candidate measures should be specified in terms of the parameters to be sampled, the rates at which they are sampled, their desired values (if any) and the transformation.
3. Extreme care is necessary in the specification of unambiguous rules for starting and stopping measurement.
4. It is necessary to conduct measurement selection empirical studies to collect data on the candidate measures during training for subsequent measure selection analyses.
5. Testing means of individual measures for significant changes between early and late in training reduces measurement; however, this method does not consider the complexity of performance, the inter-relations between measures and does not provide a method to weight measures for the construction of an overall score.
6. Eliminating highly correlated measures is an effective method to reduce redundant information, serves as a first step filter and permits the analyst a little latitude to specify extra measures in selected areas of uncertainty; also, it is necessary if multivariate analyses are to be used.
7. Canonical correlation analyses are effective for selecting those measures out of a battery that predict later measures; however, the method (a) often produced asymmetrical predictive and criterion sets, (b) was

difficult to interpret and reduce to an algorithm required for mapping measurement into an adaptive logic, and (c) was omitted from further development at this time. However, it may be useful for diagnosis and prescription of performance in more complex, or branching adaptive logics.

8. The multiple discriminant analysis can be modified to form an effective technique for selecting and weighting those measures which best discriminate between early and later training; however, in order to use this method, it is necessary to:
 - a. collect data on all major tasks and variations to those tasks (such as center-of-gravity change, turbulence, etc.) both early and late in training,
 - b. remove highly correlated measures,
 - c. have a minimum of 5 to 7 times as many observations as variables (candidate measures),
 - d. correct statistically for repeated observations on the same trainees (if repeated observations were taken),
 - e. specify the minimum communality of any measure and the minimum number of measures (in terms of percent variance of the smallest factor),
 - f. stabilize the beta weights using modified "ridge" analysis techniques for more reliable prediction.
9. The methods used to partition the variance due to repeated observations and to stabilize the weighting coefficients should be further studied along with methods to reduce sampling requirements for more efficient data collection.
10. The measures and weighting coefficients that emerge from the modified multiple discriminant analysis can be used to form a single score, the discriminant function, for use by adaptive logics that require a single score.
11. Control input measures were often important in describing the differences between skilled and unskilled performance.

MEASURE IMPLEMENTATION

The recommended weights and measures which resulted from the multiple discriminant analysis were mapped into the automated training system (IFM), forming a second measurement subsystem. A third measurement subsystem was created by modifying the adaptive logic to operate on norms of the original IFM measures, based on data acquired during measure selection studies. The major conclusions of the implementation effort were:

1. Means and standard deviations of the discriminant function must be computed in measurement space (DISCRIM SELECT output is in discriminant space) to determine criterion performance.
2. The discriminant model can misclassify poor performance if that poor performance is on a negatively weighted measure that has a magnitude outside the measurement space of the original data (used to produce the discriminant function).
3. Misclassification is easily circumvented by a heirarchical algorithm which first tests negatively weighted measures to insure that they are within 4-sigma of their average in the original data. If the unweighted measure fails the test, the discriminant function is set to 2.7-sigma. If the measure passes the test, the discriminant function is computed.
4. A rational way to scale different measurement system outputs into the adaptive logic is through z-scores of criterion performance.
5. Real-time programming of the measures, weights, start and stop rules, and heirarchal model was achieved in the TRADEC/IFM within the 50 millisecond program cycle time; measurement included control input power approximation in the frequency domain through the use of digital high and low-pass filters, sampling 20 times per second.

MEASURE EVALUATION

Empirically derived measurement systems were substituted in an existing automated instrument flight maneuvers training system with the result that time to train to the same criterion was reduced 34-40%. It was concluded that:

1. If this result holds in subsequent validation, the users of advanced and retro-fitted training systems (that contain measurement improved by empirical techniques) can look forward to improved efficiency and utilization of those devices.
2. In existing automated training devices that have measurement, these levels of increased efficiency should result by modifying the adaptive logic to operate on actual performance norms rather than assumed norms in their scoring algorithms.
3. The approach taken in the development of the modified multiple discriminant analysis for selecting measures (DISCRIM SELECT) was partially validated.

4. The discriminant model measurement appeared to be sensitive to piloting technique and to provide more reliable performance feedback.
5. The discriminant model can be expected to produce better measurement in future efforts than was shown in the evaluation because it can select and properly weight (along with system measures) student history variables (such as age) which have been shown to be very important.
6. The measure production and selection techniques herein described have produced improvement to analytically derived measurement of a sufficient magnitude to warrant application of these techniques with the end goal of specifying measurement for future and existing flight training systems. In order to apply the techniques, data collection in field training environments is required.

Although the purpose of the program was to develop and evaluate measurement, several conclusions concerning the design of automated training systems are related to measurement and can be made from the data:

1. Linear, single score adaptive logics similar to the configuration of IFM may not be efficient enough to use in operational training. The interaction between the syllabus exercises, adaptive logic and measurement does not always permit the good trainee to advance rapidly. Marked improvement should result by:
 - a. Limiting the number of exercises within a maneuver to only those that have operational relevance.
 - b. Removing exercises from the main line sequence (or removing them altogether) that only provide task variation or stressors such as turbulence.
 - c. Strongly inhibit, or remove altogether, backward movement through the syllabus.
 - d. Construct the score on the basis of performance norms.
2. Adaptive logics which require a single score do not take advantage of the power of measurement to diagnose performance and lead to better prescription of training. Branching logics based on more than one measure should be more efficient.
3. It may not be necessary to expect a student to perform within 2-sigma of end of training criteria in all cases. The measurement system should be designed to evaluate

NAVTRAEQUIPCEN 74-C-0063-1

performance against norms based on time in training.
Assumed norms can be used until sufficient data accrues
to change them.

SECTION VII

RECOMMENDATIONS

It is recommended that:

1. The techniques described herein be improved and used to produce and select measurement for existing and future automated flight training systems.
2. An operational flight training site (or sites) be selected for performance data collection in a military flight training simulator environment. Subsequent analyses of the data should lead eventually to a specification for measurement for the maneuvers and class of aircraft tested.
3. Initial field measurement activities be limited to instrument flight or weapon delivery phases of simulator training where initial conditions and prescribed flight paths are known and specifiable; however, it is possible and recommended that other flight regimes (where criteria can be specified) be explored for measurement possibility.
4. The results of initial field studies (ie: recommended measures and weights) should be installed in the field systems, and an evaluation of the new measurement should be conducted to determine the training impact (similar to the Phase III evaluation reported herein).
5. Continued improvement to DISCRIM SELECT be undertaken by incorporating other nonsystem performance measures such as age, time in training, and student history, and by further research with the existing data base.
6. Consideration be given to add to the Phase I data base some early trials with turbulence and turbulence in combination with aft center-of-gravity, so that measures for those task stressors can be produced.
7. Statistical issues brought about by the use of multivariate methods for measure selection be further studied; these issues include, but are not limited to (a) methods to partition the variance due to repeated measures, (b) methods to stabilize the weighting coefficients and (c) methods to possibly reduce sampling requirements.
8. Existing and future single score, linear adaptive logics be limited as described herein, and that scoring be based on performance norms throughout training. Future systems should contain performance data files that make the conversion from initially assumed norms to actual norms convenient, and changes to the scoring algorithms possible without reprogramming.

NAVTRAEQUIPCEN 74-C-0063-1

9. Future automated training systems be designed with branching (or at least lateral) logics that make decisions on more than one performance score, and that the construction and weighting of those scores be readily amenable to change without reprogramming.
10. IFM be modified, and a study conducted to determine the efficacy of (a) a limited linear adaptive logic (as in Conclusions), (b) a lateral logic which permits graduation from task variation trials to the next maneuver, and (c) a limited criterion test, branching logic. Since the mechanisms are all in place, minimum resource expenditures could provide substantial guidance for future system designers.

REFERENCES

- Blackman, R.W. and Tukey, J. The Measurement of Power Spectra. New York, Dover, 1959.
- Charles, J.P., Johnson, R.M. and Swink, J.R. Automated Flight Training (AFT) Instrument Flight Maneuvers. NAVTRAEQUIPCEN 71-C-0205-1. U.S. Naval Training Equipment Center, Orlando, Florida, 1972.
- Cooley, N.W. and Tukey, J. An Algorithm for the Machine Calculation of Complex Fourier Series, Mathematics of Computation, Vol. 19, No. 90, April 1965, pp 297-301.
- Cooley, N.W. and Lohnes, P.R. Multivariate Data Analysis. New York: John Wiley, 1971.
- Erickson, E.S., Kapsis, P.B., Ciolkosz, M.D. Software Documentation for the Research Tool Digital Computer System Volume II Program Report. U.S. Navy, NAVTRADEVCEEN 67-C-0196-7, September 1969.
- Erickson, E.S., Kapsis, P.B., Ciolkosz, M.D. Software Documentation for the Research Tool Digital Computer System Volume IIA Detailed Program Description. U.S. Navy, NAVTRADEVCEEN 67-C-0196-7, September 1969.
- Heal, L.W. VULT-Vanderbilt Statistical Package, Fortran IV, Sigma 7 BPM/BTM, Catalog No. 890040-11V300, Vanderbilt University, Nashville, 1971.
- Hoerl, A.E. and Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, Vol. 12, No. 1, February 1970, pp 55-67.
- Hoerl, A.E. and Kennard, R.W. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, Vol. 12, No. 1, February 1970, pp 69-82.
- Johnson, R.M. AFT Program Description. Contract N61339-71-C-0205, Report SDR-111(AFT)PD, Logicon, San Diego, May 1972.
- Kapsis, P.A., et al. Software Documentation for the Research Tool Digital Computer System Volume I Math Model Report. U.S. Navy, NAVTRADEVCEEN 67-C-0196-7, September 1969.
- Knoop, P.A. and Welde, W.E. Automated Pilot Performance Assessment in the T-37: A Feasibility Study. AFHRL-TR-72-6, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, April 1973.
- Lane, N.E. The Influence of Selected Factors on Shrinkage and Overfit in Multiple Correlation. Naval Aerospace Medical Research Laboratory, Naval Aerospace Medical Institute, Pensacola, Florida, September 1971.

- Norman, D.A. Personal Communication on the Implementation of a Second-Order, Low-Pass Digital Filter Program. 1973.
- Obermayer, R.W. and Vreuls, D. Measurement for Flight Training Research. Proceedings of the 16th Annual Meeting of the Human Factors Society, Beverly Hills, California, October 1972.
- Schori, T.R. and Tindall, J.F. Multiple Discriminant Analysis: A Repeated Measures Design. Virginia Journal of Science, Vol. 23, 1972, pp 62-63.
- Searle, S.R. Linear Models. New York: John Wiley, 1971.
- Siegel, S. Nonparametric Statics for the Behavioral Sciences, New York, McGraw-Hill, 1956.
- Smode, A.F. Human Factors Inputs to the Training Device Design Process. NAVTRADEVCEEN 69-C-0298-1. U.S. Naval Training Device Center, Orlando, Florida, September 1971, pp 327-378.
- Villasensor, A.J. Digital Spectral Analysis. NASA TN D-4510. National Aeronautics and Space Administration, Greenbelt, MD June 1968.
- Vreuls, D. and Obermayer, R.W. Study of Crew Performance Measurement for High-Performance Aircraft Weapon System Training Air-to-Air Intercept. NAVTRADEVCEEN 70-C-0059-1, U.S. Naval Training Device Center, Orlando, Florida, February 1971.
- Vreuls, D. and Obermayer, R.W. Emerging Developments in Flight Training Performance Measurement. U.S. Naval Training Device Center 25th Anniversary Commemorative Technical Journal, November 1971, pp 199-210.
- Vreuls, D., Obermayer, R.W., Goldstein, I., and Lauber, J.W. Measurement of Trainee Performance in a Captive Rotary-Wing Device. NAVTRAEQUIPCEN 71-C-0194-1. U.S. Naval Training Equipment Center, Orlando, Florida, July 1973.
- Vreuls, D., Obermayer, R.W., and Goldstein, I. Trainee Performance Measurement Development Using Multivariate Measure Selection Techniques. NAVTRAEQUIPCEN 73-C-0066-1. U.S. Naval Training Equipment Center, Orlando, Florida, September 1974.
- Vreuls, D. and Goldstein, I. In Pursuit of the Fateful Few: A Method for Developing Human Performance Measures for Training Control, in NTEC/Industry Conference Proceedings. NAVTRAEQUIPCEN IH-240. U.S. Naval Training Equipment Center, Orlando, Florida, November 1974.
- Winter, B.B. Nonparametric Density Estimation and Statistical Discrimination. Psychological Bulletin, Vol. 81, No. 6, 1974, pp 371-379.

APPENDIX A

RAW DATA AND MEASUREMENT FUNCTIONS AND TRANSFORMS
AVAILABLE IN CURRENT MEASUREMENT PROGRAMS

TABLE 32. REAL TIME RAW DATA PARAMETERS FROM SIMULATOR

PARAMETER	UNITS	ABBREVIATION
1. SYSTEM CLOCK COUNT		CLOK
2. ELEVATOR STICK FORCE	POUNDS	ELVF
3. ELEVATOR STICK DISPLACEMENT	INCHES	ELVS
4. ANGLE OF ATTACK	UNITS	ALPH
5. PITCH ATTITUDE	DEGREES	PTCH
6. CLIMB/DESCENT RATE	FEET PER MINUTE	HDOT
7. ALTITUDE	FEET	ALT
8. RIGHT THROTTLE DISPLACEMENT	DEGREES	THRR
9. AIRSPEED	KNOTS	A/S
10. AILERON STICK FORCE	POUNDS	AILF
11. AILERON STICK DISPLACEMENT	INCHES	AILS
12. ROLL ATTITUDE	DEGREES	ROLL
13. TURN RATE	DEGREES PER SECOND	TURN
14. HEADING	DEGREES	HEAD
15. RUDDER PEDAL FORCE	POUNDS	RUDF
16. RUDDER PEDAL DISPLACEMENT	INCHES	PED
17. SIDESLIP	DEGREES	BETA
18. TURBULENT AIR INTENSITY	ARBITRARY UNITS	RUFF

TABLE 33. GLOSSARY OF START/STOP FUNCTIONS¹

MNEMONIC	FUNCTION	START/STOP WHEN:
B		Beginning of Record
E		End of Record
P		End, Best Fit Power of 2
G	PAR > DSR [*]	Parameter Greater than Desired Value
L	PAR < DSR	Parameter Less than Desired Value
O	PAR - DSR > TOL	Absolute value of parameter minus desired value is greater than (outside of) tolerance
I	PAR - DSR < TOL	Absolute value of parameter minus desired value is less than (inside) tolerance
CO	PAR - INIT > TOL	Absolute value of parameter minus its initial value is greater than tolerable (or the change from initial is outside of tolerance)
CI	PAR - INIT < TOL	Absolute value of parameter minus its initial value is less than the tolerance

TABLE 34. GLOSSARY OF LOGICAL OPERATORS FOR COMBINING START/STOP FUNCTIONS¹

MNEMONIC	EACH PAIR OF FUNCTIONS (F) IS EVALUATED TRUE IF:
A	F ₁ is True and F ₂ is True
O	F ₁ is True or F ₂ is True
N	F ₁ is True and F ₂ is False
R	F ₁ is False and F ₂ is False

¹These logical and relational expressions could be expanded as necessary.

TABLE 35. GLOSSARY OF TRANSFORMATIONS

MNEMONIC	TRANSFORMATION
INIT	Initial Scalar Value
FINL	Final Scalar Value
AINL	Absolute Initial Scalar Value
AFIN	Absolute Final Scalar Value
MIN	Minimum Value
MAX	Maximum Value
AVG	Average Value $\frac{1}{N} \sum_{n=1}^N x$
AAE	Average Absolute Value $\frac{1}{N} \sum_{n=1}^N x $
ERS	Error Squared Value $\frac{1}{N} \sum_{n=1}^N x^2$
VAR	Variance $\sum_{n=1}^N x^2 - \frac{1}{N} (\sum_{n=1}^N x)^2$
RMS	Root-Mean-Square $(\frac{1}{N} \sum_{n=1}^N x^2)^{1/2}$
SDV	Standard Deviation $\frac{1}{N-1} (\sum_{n=1}^N x^2 - \frac{1}{N} (\sum_{n=1}^N x)^2)^{1/2}$
TOT	Time Out of Tolerance in Seconds and Tenths
RNG	Range, Distance Between the Largest and Smallest value
ELT	Elapsed Time in Seconds and Tenths
ZRX	No. Zero Crossings per Second
AVX	No. Average Crossings per Second
AUTO	Auto Covariance Function

TABLE 35. GLOSSARY OF TRANSFORMATIONS (continued)

MNEMONIC	TRANSFORMATION
PERD	Periodicity of Auto Covariance Function, the tau shift values and covariance at peaks.
MLTR	Multiple Regression of a Parameter x and its derivative (x) on Parameter y (Cooley and Lohnes, 1962). This particular transform computes successive multiple regressions of x, x on later (tau) values of y, (as in an auto covariance function) until maximum multiple regression coefficient is found. It returns (1) Tau in seconds, (2) the coefficient of multiple regression (3) the Beta weights and (4) B-weights at the point of maximum multiple regression.
HARM	Harmonic Analysis using procedures outlined Blackman and Tukey (1959), Cooley and Tukey (1965) and Villasenor (1968) produced the power spectral density function for the requested bandwidth.
FLTR	Relative power between 2 and 6 radians per-second using a pair of low-pass second-order digital filters as described by Norman (1973).

APPENDIX B

CANDIDATE MEASURE MEANS AND t-TESTS BY MANEUVER

TABLE 36. AVERAGE MANEUVER 1 (STRAIGHT & LEVEL) MEASURES

MEASURE	NO STRESS					AFT CG		TURB	
	DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	CG1 DAY8	CG3 DAY9
ELRG	1.73*	.97*	.83	.82	1.39*	1.03*	.88	1.24*	1.47*
ELF1	.03*	.01	.01	.01	.07*	.06	.05	.02*	.08*
ELF2	.72	.68	.66	.64	.27*	.27*	.23	.78*	.32*
AIRG	1.20*	.62	.58	.56	.90*	.65	.63	.92*	1.00*
AIF1	.06*	.02	.03	.03	.03*	.02	.02	.04*	.03
AIF2	.27*	.23*	.19	.19	.26	.27*	.24	.25*	.28*
PDRG	.15*	.08	.07	.07	.12*	.06	.07	.10*	.09
PDF1	.01	.01	.01	.01	.01	.01	.01	.01	.01
PDF2	.14*	.12	.11	.11	.12	.14*	.11	.13*	.12
ALRG	2.46*	1.40*	1.18	1.20	3.66*	2.76*	2.37	1.86*	3.84*
ALSD	.46*	.26*	.21	.21	.68*	.52*	.44	.32*	.72*
PTRG	4.15*	2.13*	1.73	1.68	4.71*	3.43*	2.94	2.24*	4.36*
PTSD	.90*	.45*	.36	.32	.96*	.70*	.58	.43*	.84*
ROAA	2.33*	1.74*	1.71	1.50	2.42*	1.88	1.62	1.52	1.66
RORM	3.17*	2.21*	2.25	1.89	3.18*	2.38	2.06	1.97	2.13
PSRM	2.89*	1.85*	1.73	1.52	2.32	1.84	1.63	1.21*	1.33
PSRG	4.50*	2.84*	2.69	2.47	4.06*	3.03*	2.56	2.44	2.72
HAA	.06*	.03*	.02*	.01	.05*	.03*	.02	.01	.02*
HRG	.19*	.09*	.07*	.05	.16*	.10*	.08	.05	.08*
HDAA	.49*	.23*	.18*	.15	.42*	.29*	.23	.17	.27*
HDRG	2.51*	1.25*	.99	.89	2.27*	1.62*	1.31	1.04	1.64*
ASAA	6.99*	4.07*	2.83	2.75*	4.75*	4.16*	3.17	2.64	2.87
ASRG	15.10*	10.68*	6.89	7.10	10.99*	9.51*	6.91	7.34	8.37*

* Measure significantly different than Day 6 (vs Days 2 and 4) and Day 7 (vs Remaining days), $p < .05$ based on t-tests; 142 D/F.

TABLE 37. AVERAGE MANEUVER 2 (CLIMBS & DESCENTS) MEASURES

MEASURE	NO STRESS							AFT CG			TURB	
	DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	CG1	DAY8	CG3	DAY9	
ELF1	.04*	.02	.02	.02	.10	.11	.10	.02	.16*			
ELF2	.30*	.24*	.23	.22	.20*	.17*	.14	.33*	.21			
ALRG	1.95*	1.27	1.20	1.20	3.38*	2.84*	2.30	1.74*	3.60*			
ALSD	.34*	.20*	.19	.18	.58*	.47*	.39	.26*	.63*			
PTSD	.58*	.32*	.29	.28	.80*	.61*	.48	.32*	.70*			
HDA	.35*	.19*	.17	.16	.42*	.28*	.22	.15	.24*			
AIF1	.04*	.02*	.03	.03	.03	.03	.02	.04*	.02			
AIF2	.25*	.20*	.18	.17	.24*	.23*	.20	.22*	.24*			
ROA	2.73*	1.92*	1.72	1.49	2.59*	2.03*	1.75	1.35	1.65			
ROR	3.67*	2.52*	2.28*	1.89	3.34*	2.59*	2.19	1.70	2.10			
PDF1	.01	.01	.01	.01	.01*	.01	.01	.01	.01			
PDF2	.15*	.13*	.12	.11	.13*	.13*	.12	.13*	.13*			
PSA	2.62*	1.85*	1.56	1.50	2.13*	1.91*	1.44	.89*	1.06*			
PSR	2.88*	2.02*	1.71	1.66	2.39*	2.12*	1.60	.98*	1.19*			
TUR	3.51*	1.72*	1.56	1.53	2.46*	1.72*	1.46	2.09*	2.12*			
TUA	2.59*	1.27*	1.14	1.08	1.79*	1.30*	1.07	1.46*	1.52*			
BER	.62*	.52	.47	.48	.58*	.55	.50	.54*	.56*			
ASA	6.49*	4.65*	3.67	3.80	6.30*	4.76*	3.59	3.19*	3.20*			
THR	2.54*	1.96	1.70	1.86	2.88*	2.35*	1.83	1.65	2.41			

* Measure Significantly different than Day 6 (vs Days 2 and 4) and Day 7 (vs Remaining Days), $p < .05$ based on t-test; 142 D/F.

TABLE 38. AVERAGE MANEUVER 3 (LEVEL TURNS) MEASURES

MEASURE	NO STRESS					AFT CG			TURB	
	DAY1	DAY3	DAY5	DAY7		DAY2	DAY4	DAY6	CG1 DAY8	CG3 DAY9
ELF1	.02*	.02	.02	.02		.05	.04	.05	.02*	.06*
ELF2	.75*	.70*	.66	.63		.45*	.45*	.40	.77*	.48*
ALRG	2.38*	1.63	1.59	1.63		4.08*	3.38	3.41	2.19*	4.74*
ALSD	.51*	.33	.32	.32		.86*	.71	.70	.41*	.95*
PTSD	1.05*	.57*	.50	.49		1.19*	.91	.84	.53	1.02*
AIF1	.05*	.02	.03	.03		.03	.03	.03	.05*	.04*
AIF2	.35*	.31*	.28	.26		.36	.37*	.33	.31*	.35*
ROAA	5.11	4.70	4.24	4.24		4.78*	5.20	4.13	3.18*	3.87
PDF1	.02	.01	.02	.01		.02	.02	.02	.02	.02
PDF2	.17*	.16*	.13	.12		.16*	.17*	.14	.14	.13
BERG	1.21	.91*	.99*	1.12		1.16*	1.07	1.01	1.36*	1.43*
BERM	.69*	.65*	.56	.56		.68	.68	.63	.61	.62
ASAA	6.69*	3.34*	2.69	2.77		5.67*	4.03*	3.06	2.47	3.09
ASRM	7.58*	3.75*	3.06	3.15		6.37*	4.55*	3.51	2.81	3.51
HAA	.10*	.04*	.02	.02		.08*	.04	.04	.02	.03*
THRG	1.80*	1.10	.81*	1.16		1.67*	1.64*	1.14	1.09	1.68*

* Measure significantly different than Day 6 (vs Days 2 and 4) and Day 7 (vs remaining days), based on t-test; 142 D/F.

TABLE 39. AVERAGE MANEUVER 4, SEGMENT 2 (INITIAL CLIMB/DIVE TURN) MEASURES

MEASURE	NO STRESS							AFT CG			TURB	
	DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	CG1		CG3		
								DAY8	DAY9			
ELF1	.01*	.01	.01	.01	.01	.01	.01	.01	.01	.02*		
ELF2	1.68*	1.66*	1.63	1.59	.75	.77*	.71	1.77*	1.77*	.78*		
ALRG	3.43*	2.40	2.36	2.31	5.10*	4.62*	4.11	2.88*	2.88*	5.81*		
HDA	.44*	.24*	.21	.21	.45*	.33*	.27	.21	.21	.33*		
THRG	3.11*	1.94	1.64	1.79	3.03*	2.37	2.12	1.87	1.87	2.78*		
ASAA	9.07*	4.31*	3.50	3.43	5.56*	4.30*	3.58	3.11	3.11	5.09		
AIF1	.03	.02	.02	.02	.02	.01	.01	.04*	.04*	.03		
AIF2	2.36	.37	.34	.33	.44	.46*	.40	.36	.36	2.34		
BERM	3.16	.65	.58	.60	.61	.61	.55	.59	.59	1.40		
ROAA	6.18	3.08	2.81	3.05	4.45*	3.55	3.38	2.76*	2.76*	4.98		
PDF1	.01	.01	.01*	.01	.01	.01	.01	.01*	.01*	.01*		
PDF2	.30	.17*	.15	.15	.17*	.18*	.15	.15	.15	.27		
PSAF	7.22	3.67*	4.20	5.27	5.06	4.43	4.06	4.97	4.97	6.16		
TIME	62.40	62.82	62.83	63.16	62.34	62.78	62.50	62.03*	62.03*	62.26		

*Measure significantly different than Day 6 (vs Days 2 and 4) and Day 7 (vs remaining days), $p < .05$ based on t-tests, 142 D/F.

TABLE 40. AVERAGE MANEUVER 4, SEGMENT 3 (CLIMB/DIVE & TURN REVERSAL) MEASURES

MEASURE	NO STRESS							AFT CG		TURB	
	DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	CG1	CG3	DAY8	DAY9
ELF1	.17	.19	.25	.20	.12*	.19	.23	.23	.17	.23	.17
ELF2	1.58	1.53	1.52	1.46	.74	.73	.69	1.68*	.85*	1.68*	.85*
ALRG	2.98*	2.01	1.79	1.96	5.46*	4.14	3.89	2.56*	5.95*	2.56*	5.95*
HDAF	1.04*	.66	.73	.64	.86	.73	.78	.67	.62	.67	.62
AIF1	.19	.19	.26	.20	.10*	.16	.20	.23	.13*	.23	.13*
AIF2	2.51	.54	.55	.59	.60	.61	.60	.75*	2.61	.75*	2.61
BERG	1.44	1.04*	.98*	1.25	1.38*	1.15	1.15	1.66*	1.88*	1.66*	1.88*
ROAF	16.86	15.10	19.35	16.16	9.85*	14.02	16.44	16.44	14.47	16.44	14.47
TIME	22.23*	16.35*	12.33	12.63	26.30*	19.90*	15.91	10.18*	15.39*	10.18*	15.39*
PDF1	.16	.18	.24	.19	.08*	.14	.19	.21	.11*	.21	.11*
PDF2	.33	.18	.15	.17	.18	.19*	.16	.17	.28	.17	.28

* Measure Significantly different than Day6 (vs Day 2 and 4) and Day 7 (vs Remaining Days), $p < .05$ based on t-test, 142 D/F.

TABLE 41 . AVERAGE MANEUVER 4, SEGMENT 4 (FINAL CLIMB/DIVE TURN) MEASURES

MEASURE	NO STRESS						AFT CG			TURB	
	DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	CG1	CG3	DAY8	DAY9
ZLF1	.01*	.01	.01	.01	.03*	.02	.02	.01*	.03*		
ELF2	1.65	1.64	1.64	1.58	.75	.76	.71	1.76*	.78*		
ALRG	4.21*	2.87	2.73	2.81	6.98*	5.69	5.44	3.85*	7.77*		
HDAA	.72*	.41*	.37	.34	.70*	.54*	.45	.32	.49*		
THRG	5.03*	4.17	3.66	3.93	6.08*	5.71	5.43	4.36	7.33*		
ASAA	9.47*	6.28*	4.37	4.18	9.69*	7.20*	5.50	4.27	5.29*		
AIF1	.03*	.02	.03	.02	.02	.02	.02	.04*	.03*		
AIF2	.46*	.41	.39	.38	.49	.51	.47	.41	.47*		
BERM	.75	.72	.65	.70	.68	.70	.66	.68	.67		
ROAA	9.35*	7.94*	7.68	7.24	9.91*	8.37*	7.88	6.39*	7.67		
PDF1	.01	.00*	.01	.01	.01	.00	.00	.01	.01		
PDF2	.18*	.18	.15	.16	.17	.19*	.16	.15	.15		
PSAF	15.27*	8.64*	7.45	5.77	15.41*	11.12	8.79	6.51	13.27*		
TIME	81.78*	79.04*	76.46	76.99	82.68*	82.47*	79.33	74.44*	79.08		

* Measure significantly different than Day 6 (vs Day 2 and 4) and Day 7 (vs Remaining Days), $p < .05$ based on t-test, 142 D/F.

NAVTRAEQUIPCEN 74-C-0063-1

APPENDIX C

EQUIVALENT MEASURES BY MANEUVER

TABLE 42. MANEUVER 1 (STRAIGHT & LEVEL) EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELRG	1 *	1	1	1
ELF1				
ELF2				
AIRG				
AIF1				
AIF2	2	2	2	2
PDRG				
PDF1				
PDF2	2	2	2	2
ALRG	1	1	1	1
ALSD	1	1	1	1
PTRG	1	1	1	1
PTSD	1	1	1	1
ROAA		4		
RORM		3	3	3
PSRM		3		3
PSRG				
HAA				
HRG		4	4	4
HDAA		4	4	4
HDRG		4	4	4
ASAA				
ASRG				

* Chains of measures which intercorrelate, $r_{.90}$, for each comparison day.

TABLE 43. MANEUVER 2 (CLIMBS & DIVES) EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELF1				
ELF2				
ALRG	1	1	1	1
ALSD	1	1	1	1
PTSD	2	2	2	1
HDAA	2	2	2	1
AIF1				
AIF2	3	3	3	3
ROAA	4	4	4	4
RORM	4	4	4	4
PDF1				
PDF2	3	3	3	3
PSAA		5	5	5
PSRM		5	5	5
TURM		6	6	6
TUAA	6	3	3	3
BERM	6	3	3	3
ASAA				
THRG				

* Chains of measures which intercorrelate, $r \geq .90$, for each comparison day.

TABLE 44. MANEUVER 3 (LEVEL TURNS) EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELF1				
ELF2				
ALRG	1 *	1	1	1
ALSD	1	1	1	1
PTSD				1
AIF1				1
AIF2				1
ROAA				
RORM				
PDF1				
PDF2	3		3	3
BERG				
BERM	3		3	3
ASAA				
ASRM	4	4	4	4
HAA	4			
THRG				

* Chains of measures which intercorrelate, $r \geq .90$, for each comparison day.

TABLE 45. MANEUVER 4, SEGMENT 2 (INITIAL CLIMB/DIVE TURN)
EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELF1				
ELF2				
ALRG				
HDAA				
THRG				
ASAA	1 *			1
AIF1	1			1
AIF2	1	2	2	1
BERM	1	2	2	1
ROAA	1			1
PDF1				
PDF2	1			1
PSAF	1			
TIME				

TABLE 46. MANEUVER 4, SEGMENT 4 (FINAL CLIMB/DIVE TURN)
EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELF1	1 *			
ELF2				
ALRG				
HDAA				
THRG				
ASAA				
AIF1				
AIF2	2			
BERM	2	3		
ROAA			3	3
PDF1	1			
PDF2		3		
PSAF			3	
TIME				

* Chains of measures which intercorrelate, $r \geq .90$, for each comparison day.

TABLE 47. MANEUVER 4, SEGMENT 3 (CLIMB/DIVE TURN REVERSAL)
EQUIVALENT MEASURES

CANDIDATE MEASURES	DAY 1 vs DAY 7	DAY 2 vs DAY 6	DAY 7 vs DAY 8	DAY 7 vs DAY 9
ELF1	1 *	1	1	1
ELF2				
ALRG				
HDAF				
AIF1	1	1	1	1
AIF2	2			2
BERG				
ROAF	1	1	1	1
PDF1	1	1	1	1
PDF2	2			2
TIME				

* Chains of measures which intercorrelate, $r \geq .90$, for each comparison day.

APPENDIX D

IFM Program Modifications to Incorporate Performance
Measurement Techniques

The following program changes were made to the Instrument Flight Maneuvers program to incorporate real-time performance measurement. Modifications are listed by module name whenever a new module was added, an old module deleted or the initial module was altered.

1. ATE System Parameters (APAM). This data module was changed to reflect:
 - a. The deletion of the IDIIOM graphics display buffers and associated parameters.
 - b. The deletion of data not specifically required for the performance measurement update.
 - c. The addition of data and parameters needed to support the performance measurement update.
 - d. Modifications to the Task Description Table Definition List to tailor the tasks to the IFM-PM syllabus.
 - e. The addition of the magnetic tape buffer and the associated data parameters necessary to support the magnetic tape output records.
 - f. The inclusion of the parameters and allied data required to implement the three (3) scoring modes:
 - (1) Original IFM (with turbulence removed)
 - (2) Original IFM modified to utilize Normative Data
 - (3) Discriminate Analysis
 - g. The revision and update of line printer messages, scoring tables, adaptive logic constants, boundary limits, etc., necessary to support the scoring modes and magnetic tape module.
 - h. The revision of the Difficulty Level tables to remove turbulence as a difficulty factor from the IFM runs.
2. Task Description Parameters (TDP). This data module was changed to reflect:
 - a. The deletion of tasks not required for the IFM-PM update.

- b. A change to the Task Description Table format to permit the addition of Segment Description Tables.
- c. The incorporation of Segment Description Tables (SDT's) which provide the program the following data for each maneuver, configuration and parameter.
 - (1) Rate at which parameter is sampled.
 - (2) A pointer to the Start Measurement Conditions (SMC) table.
 - (3) A pointer to the Stop Measurement Conditions (TMC) table.
 - (4) A pointer to the parameter to be measured.
 - (5) A pointer to the desired value of the parameter.
 - (5) A list of the transforms to be performed on the parameter.
- d. The incorporation of Start/Stop Measurement Conditions Tables (SMC's/TMC's). These tables describe under what conditions the measurement of each parameter listed in the SDT's is to be started and terminated. These tables contain:
 - (1) The status of the parameter.
 - (2) A pointer to the parameter to be tested.
 - (3) The function (i.e., greater than, equal to, etc., to some desired value).
 - (4) A pointer to the desired value of the parameter.
 - (5) A tolerance for the desired value.
 - (6) A conditional which generates another set of items (1) through (5) above. Examples of conditionals are: Logical OR, Logical AND, Sequential AND.
- e. The incorporation of the following real-time tables and buffers to support the performance measurement functions:
 - (1) Segment Rate Table - reflects the rate at which each parameter is to be sampled.
 - (2) Segment Description Table - a pointer which corresponds to each item in the SRT pointing to the appropriate buffer in the Event Segment Table (EST).

(3) Event Segment Table (EST) - A real-time buffer containing data for all the parameters to be sampled for the current event. It is compiled from data supplied by the SDT's for the segments required. It contains the following information:

- (a) Sampling rate.
- (b) The SMC/TMC index.
- (c) A pointer to the parameter being sampled.
- (d) A pointer to the desired value.
- (e) A list of transforms to be performed on the parameter.
- (f) A pointer to a collection buffer assigned each transform.

(4) Start Measurement/Terminate Measurement Table pointers. These are tables which point to the appropriate Start/Stop Measurement Conditions Tables. An index to these tables is placed in (3)(b) above.

(5) Collection Buffers - These buffers are used by each parameter transform to collect data in real-time. Their individual length is dependent upon the type of transform (amount of data required for the transform).

f. The addition of a table which specifies which parameters are available for output to magnetic tape.

- 3. ATE Modifications (AMOD). The emergency procedures were deleted from this module.
- 4. AFT Modifications (AFTM). No changes.
- 5. ATE Executive Routines (ATEX). The average rate of climb and rate of turn computations were removed from foreground processing and placed in the background program Parameter Update (PMUP). A routine needed to convert turn rate from radians per second to degrees per second was added.
- 6. Trim Aircraft (TRCZ). No changes.
- 7. Pseudo-Hearing (PSH). No changes.
- 8. Timing Control (TIMR). The graphics display timer was removed.
- 9. PM SDT Processor (SDT). This was a new module added to the list of foreground processors. This routine interrogated the Segment Rate Tables (SRT) and if time to sample, it fetches the appropriate parameter, performs the specified transforms and places the intermediate results in the collection buffers.

10. Input/Output (IO). The following changes were made:
 - a. Deletions to eliminate the residual GCA and IDIOM display functions.
 - b. The addition of a "\$WEOD" teletype input command which outputs an end-of-file record to the magnetic tape.
11. Parameter Update (PUP). A new background module to compute heading, bank angle, roll rate and pitch angle which was previously accomplished in the foreground mode.
12. Exercise Scheduler (EXSC). Eliminated GCA and Emergency Procedures routing. Added Coding required to save data needed for PMDP routine.
13. Exercise Terminator (EXTR). Eliminated GCA and Emergency Procedures Routing and the logic to terminate the session automatically.
14. Post Run Router (PRR). Eliminated GCA and Emergency Procedures Routing.
15. IFM Initialize (IFIN). Eliminated DR\$3 bypass routine.
16. IFM Preflight Check (PREF). No changes.
17. Controlled Take-Off (CTO). No changes.
18. Control to Basic IFM Configuration (CIFC). No changes.
19. IFM Task Selector (IFTS). This module was modified to reflect the following:
 - a. Eliminated the graphics display set-up.
 - b. Incorporated the provision for processing the SDT's and setting up the appropriate real-time tables and buffers for the selected measuring segment.
 - c. Added the option for a "Leg Complete" cognitronics message on designated legs.
 - d. Provided for following discrete lamps in the event of cognitronics failure.
 - (1) Take Control.
 - (2) Place Speed Brake In.
 - (3) Leg Complete.
 - (4) Stop Controlling Aircraft.

(5) Good Run.

- e. Incorporated a subroutine STPROC which is called from the GPM module. The purpose of this subroutine is to process the SMT/TMT tables and test the associated SMC/TMC's to determine if measurement on the corresponding segment is to start or terminate.
- f. Provided coding to allocate storage for saving the Absolute Average Errors generated for Heading, Roll and Turn Rate in the original IFM program.
- g. Initialization of the magnetic tape output buffer (MTBUFFER).

20. General Performance Monitor (GPM). This module was modified to incorporate the following changes:

- a. Provide a computation of the absolute heading and altitude differences.
- b. Open outer limits on all parameters to prevent the run from premature termination.
- c. Provide linkages for the Performance Measurement real-time modules.
- d. Compute and save the Absolute Average errors for Heading, Roll and Turn Rate for IFM magnetic tape output.

21. IFM Display List Update (IDII). This module was deleted for the Performance Measurement program.

22. IFM Data Processing (IDP). The following changes were made to this module:

- a. The ability to read the console sense switches was incorporated. Sense switches incorporated and their meanings are:

<u>Switch #</u>	<u>Meaning</u>
1	Use IFM Original scoring
2	Use Discriminate scoring
3	Use IFM Normative Data scoring

- b. Provide linkage for PM data processing module.
- c. Provide maneuver and scoring data for the magnetic tape output buffer (MTBUFFER). Sort, process and store all parameters, transforms, student file data, etc., collected by the PMDP module for end-of-run output to magnetic tape.

- d. Provides the linkage to call the magnetic tape output routine (MTOUT:1) at the completion of each run.
23. Performance Measurement Data Processing (PMDP). This was a new module added to provide for line printer and magnetic tape output of the performance measurement parameters. For each segment it lists:
- a. The maneuver.
 - b. The parameters measured along with:
 - (1) The desired value of the parameter.
 - (2) The transforms performed.
 - (3) The raw value of the transform.
 - (4) The weighting factor of the transform.
 - (5) The weighted value of the transform.
 - c. The mean and standard deviation of the total sample score.
 - d. The weighted score and adjusted weighted score for each segment and the total exercise.
 - e. The scoring mode (IFM Original, IFM Normative or Discriminate).
 - f. The adaptive logic increment selected (dependent upon the scoring mode).

The following features were also incorporated:

- a. For Discriminate scoring an upper limit was placed on parameters which have negative weighting factors. If this limit was exceeded by the raw measure value, maximum adjusted weighted score of 2.7 was used.
 - b. The set-up of the magnetic tape buffer (MTBUFFER) for segment dependent parameters (pointers, weights, raw values, weighted values, means, standard deviations, etc.).
 - c. A subroutine (PMTRAN) that transfers all performance measurement data generated in the foreground SDT:1 module to a working buffer to be processed by the PMDP background module.
24. IFM Adaptive Logic (IAL). This module was modified to permit the adaptive logic to operate on the original IFM score, the IFM Normative score or the Discriminate score depending upon

the setting of the console sense switches (see 22 above).

25. PM Parameters (PPAM). This is a new module which contains the WFIND subroutine and PM associated data tables. The WFIND subroutine locates the appropriate weighting factor as specified by the maneuver, configuration, parameter and transform. WFTAB is a table of weighting factors for these factors. MCTAB is a table of means and standard deviations for the maneuver and configuration.
26. AFT Subroutines (ASUB). This module was expanded to include a floating point to fixed number conversion (FIX), a hexadecimal to ASCII conversion (HEXASC) and an EBCDIC number to hexadecimal (BCDTOHEX).
27. Cognitronics Message Processor (COG). This module was altered to permit bypassing of the cognitronics output in the event of a hardware failure.
28. Convert Floating Point to Cognitronics Addresses (CADR). No change.
29. Data Recording (DREC). This is a new module that outputs the magnetic tape buffer (MTBUFFER) as one physical record to alternate tape units 80 and 81.

NAVTRAEQUIPCEN 74-C-0063-1

APPENDIX E

TYPICAL TRAINING PROFILES

Training profiles of typical students with each of the three measurement subsystems are shown in figures 2 - 4, on the following pages.

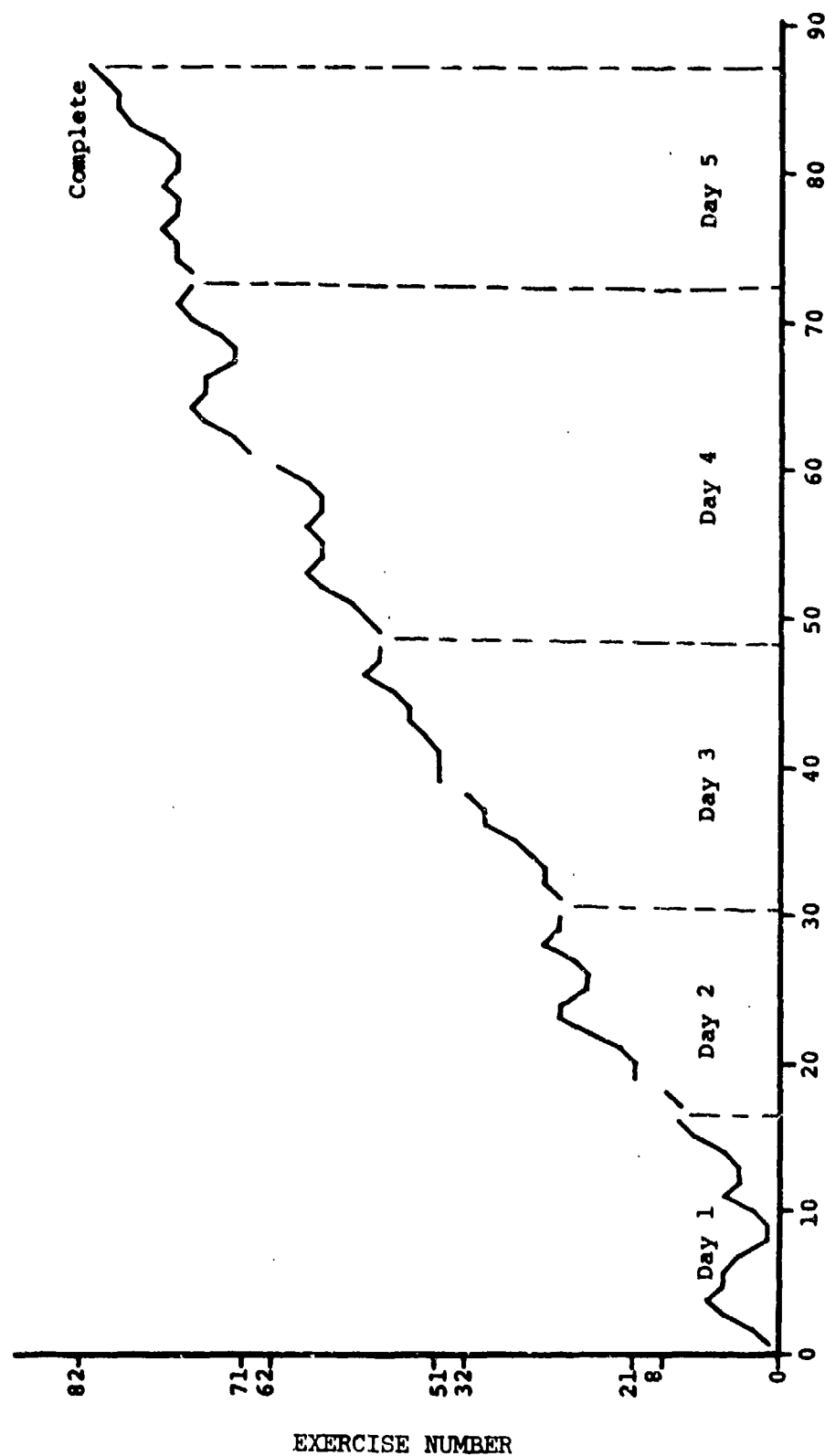


Figure 2. Typical Group 1 (Old IFM) Subject Performance

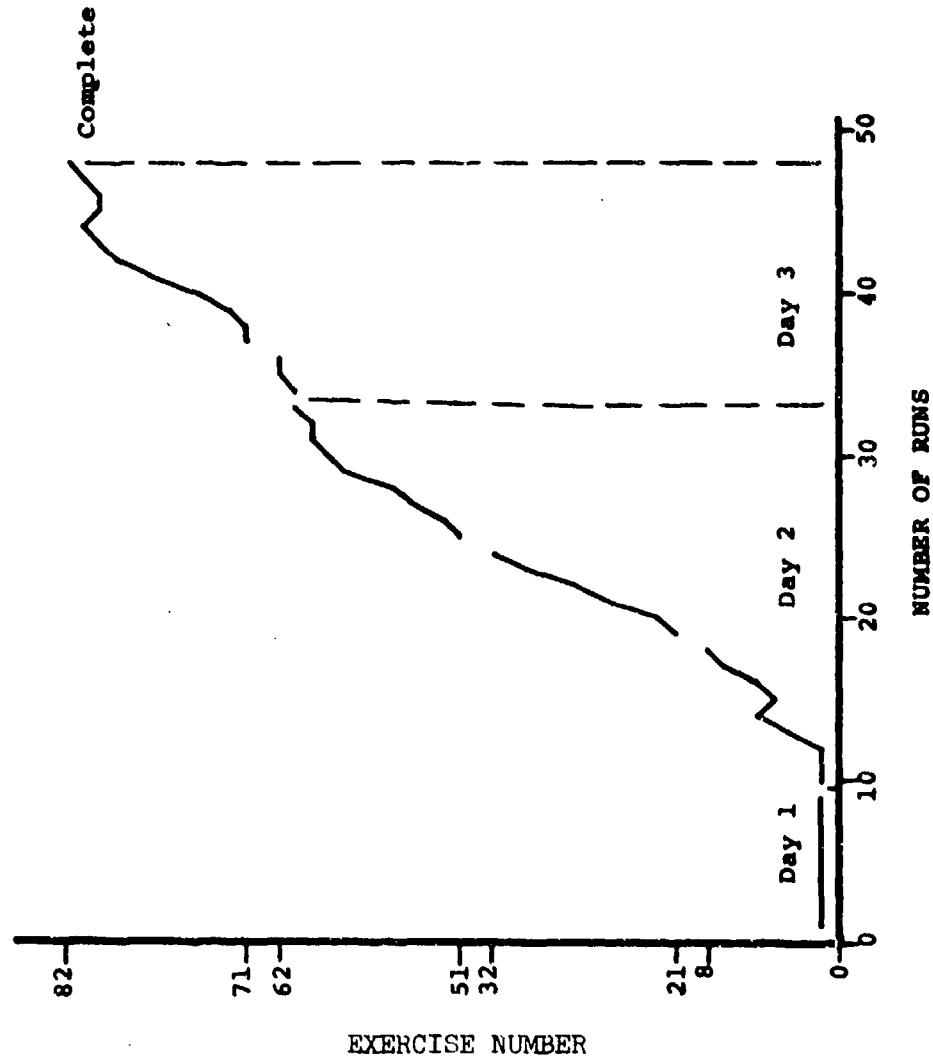


Figure 3. Typical Group II (Discrim) Subject Performance

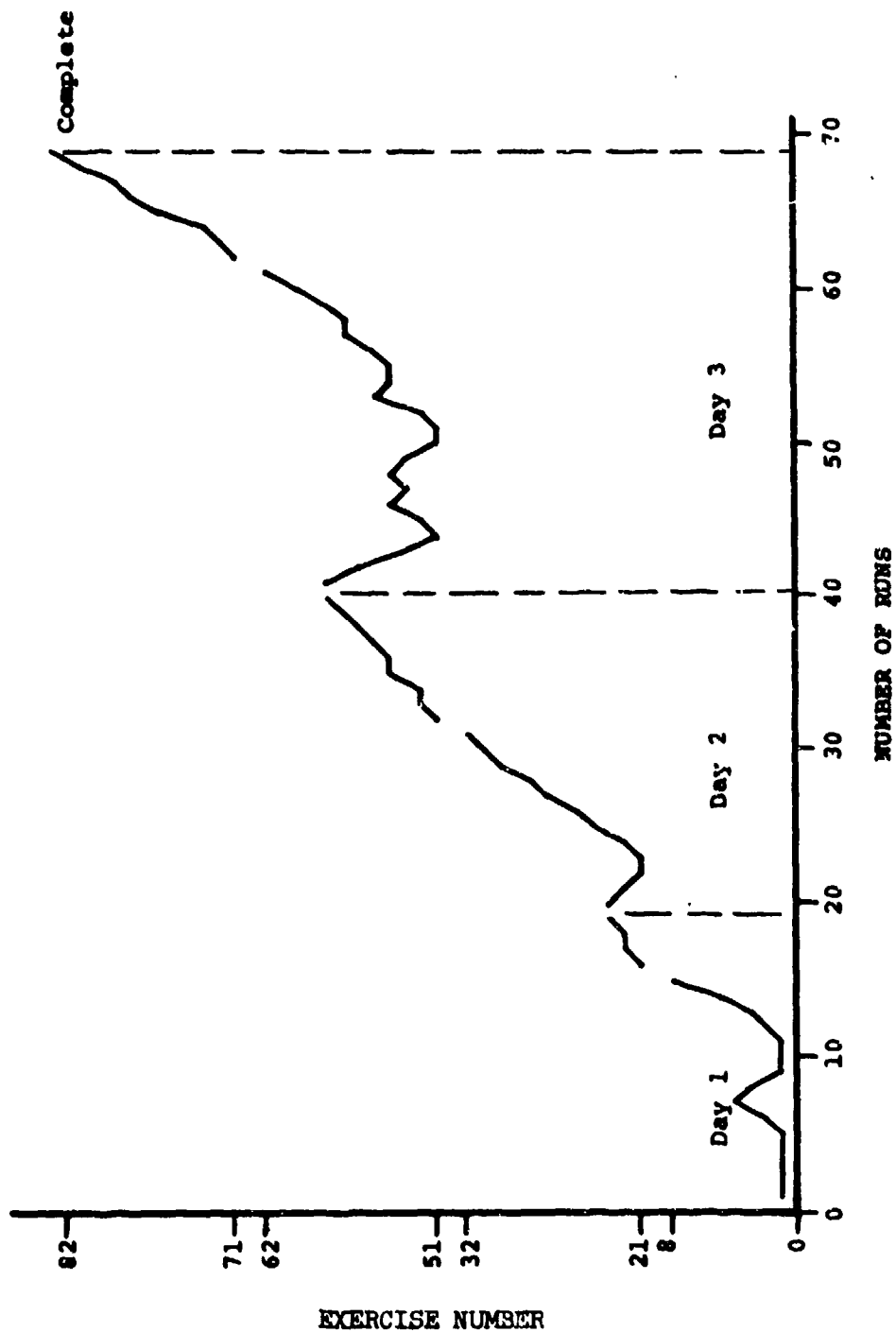


Figure 4. Typical Group III (NORM IFM) Subject Performance

APPENDIX F

COMMENTS ON AUTOMATED TRAINING SYSTEM DESIGN

Although the purpose of our work was to develop and evaluate measurement, several comments on the design of automated training systems can be made on the basis of the training problems that were observed. These comments might be helpful to designers of next generation systems.

IFM was designed under some inherent constraints that may have prevented it from being a good instructor. For example, when a pilot needed help most during early training, the coaching messages lagged too far behind his performance to be of value. There was no priority system to evaluate and correct the most important errors first. Neither were there any judgments about the reasonability of performance taking subject experience into account. Also, it was not possible to construct voice coaching on piloting technique and finesse to the extent that a good instructor would.

The syllabus and adaptive logic design of IFM may not lead to efficient training. Recall the system design (Section IV). The system required the student to master each exercise to end of training proficiency levels before advancement to the next exercise. There were many exercises within a maneuver that were composed of task variations, ordered with increasing "difficulty." The adaptive logic only moved the trainee up or down this list of exercises.

This kind of adaptive logic produced at least two problems related to inefficiency. First, when a student encountered an exercise that was difficult and performed poorly, the adaptive logic often set him back to an exercise he had already passed. But, because it had set him back, he often had to perform several trials on *exercises he had already passed* before he could try again the problem exercise. Secondly, there were too many exercises contained within each maneuver, tending to force the good pilot to perform unnecessary trials.

We are not convinced that adaptive logics that require movement through a syllabus based on a single score produce the most efficient training. Performance is multidimensional, and measurement can be made to diagnose at least major problems. For example, if a student during climbing and diving turns has problems controlling the turn, that problem is easily measured, and the logic might branch the student to a level turn exercise to at least check his ability to handle level turns,

The net result of this adaptive logic, which we shall call linear single score, is that it will very likely lead to automated training systems that increase the time required to train in operational settings over the more traditional methods.

This kind of result would be very unfortunate because the problem is not the concept of automated training, but the proper use of automated systems and the design of adaptive logics, syllabus exercises and measurement systems.

Certainly the cost and utility of producing very smart automated systems should be strongly considered during system definition. To make a system as smart as a good instructor for early in training might be very complex indeed. Once a student has an initial grasp of technique, however, automated systems might have good utility for presenting a variety of problems for practice, skill improvement, test administration and performance assessment. This utilization might represent a reasonable cost trade in terms of system complexity, would unburden the instructor of the routine, so that he could concentrate on early training and student problems, and would provide a convenient system for performance measurement and assessment.

Where there are special technique problems, such as learning the proper skill to control vehicles in unstable regimes, separate subsystems may be designed to specifically address the teaching of technique alone. Continuous adaptation of vehicle characteristics may have application in this area.

Improved performance of linear single score adaptive logic might result if backward movement through the syllabus was inhibited or eliminated altogether, and if the number of exercises within a maneuver were limited. Syllabus construction requires a great deal of care and operational input. The composition of exercises should be related to tasks which must be trained. The addition of exercises which create only task variation should not slow down training; these exercises might be considered "lateral" to the main line, and successful performance on them should cause graduation to the next "main line" exercise.

Adaptive logics can be constructed to make judgments based on performance norms relative to the student's time in training or experience level. Early in training a student may not need to perform within 2-sigma of end of course criteria. If the student is within the performance range of other students of his experience (and those norms converge on terminal criteria), then the student is performing as expected, and should advance. Systems can be designed to start operation with assumed norms that can be adjusted after sufficient data are accrued.

Branching logics may have utility where performance is expressed by more than one score. For example, a small set of criterion test exercises can be constructed. Failure to pass those exercises would result in branching to either task variation exercises or remediation exercises. If successfully passed, remediation exercises should point to the last attempted criterion exercise, but task variation exercises should point to the next criterion exercise.

HUMAN FACTORS DISTRIBUTION LIST

NOTE

Mailing labels are prepared, when needed, as a computer listing, the source of which is updated on a weekly basis. It is not practical to prepare distribution lists each time labels are prepared. Therefore such lists are prepared semiannually, and a slight discrepancy may exist between the addressees on this list and those appearing on the labels used to distribute this publication.

HUMRRO Central Division, Suite 400 Plaza Bldg, Pace Blvd at Fairfield,
Pensacola, FL 32505

USA Aeo Medical Research Lab, ATTN: Dr M. A. Hofman, P O Box
577, Ft Rucker, AL 36360

Director Human Engineering Lab, USA Aberdeen Research Development
Center, ATTN: Dr John W. Weisz, Aberdeen Proving Grounds, MD
21005

HQ, USA Training and Doctrine Command, ATTNG-CTS, Ft Monmouth,
NJ 07703

Commandant, USA Field Artillery School, Target Acquisition Dept, ATTN:
Eugene C. Rogers, Ft Sill, OK 73503

Director Human Relations Research Organization, 300 N Washington St,
Alexandria, VA 22314

Human Relations Research Organization, Division No. 1, Systems Operation,
300 N Washington St., Alexandria, VA 22314

USA Research Institute Behavior, Social Sciences, 1300 Wilson Blvd,
Arlington, VA 22209

Chief Research Office, Office Deputy Chief of Staff for Personnel, Dept
of Army, Washington, DC 20310

Asst Secretary Navy, R-D, Dept of Navy, ATTN: Dr S. Koslov 4E741,
Washington, DC 20350

Chief Naval Research, Code 458, Dept of Navy, Arlington, VA 22217

HUMAN FACTORS DISTRIBUTION LIST (CONT)

Chief Naval Research, Psychological Sciences, Code 450, Dept of Navy,
Arlington, VA 22217

Chief Naval Operations, OP-14C, Dept of Navy, ATTN: M. K. Malenorn,
Washington, DC 20350

Chief Naval Operations, OP-07T16, Dept of Navy, ATTN: Dr J. J. Collins,
Washington, DC 20350

Chief Naval Operations, MSC OP-701E2, Dept of Navy, ATTN: CDR H. J.
Connery, Washington, DC 20350

Chief Naval Material, MAT 031M, Washington, DC 20360

Chief Naval Material, 03424, C P 6, Room 820, Dept of Navy, ATTN:
Arnold L. Rubinstein, Washington, DC 20360

Bureau Naval Personnel, ATTN: PERS A3, Arlington Annex, Washington,
DC 20370

Commandant of Marine Corps, Code AO3C, Washington, DC 20380

Director, Defense Research Engineering, ATTN: LCOL H. Taylor,
OAD R-D, Washington, DC 20301

ERIC Clearinghouse on Educational Media (Technical), Stanford University,
Stanford, CA 94305

Grumman Aerospace Corp, Plant 47, ATTN: Mr Sam Campbell, Bethpage,
LI, NY 11714

Texas Technical University, Psychology Dept, Box 4100, ATTN: Dr
Charles Holcomb, Lubbock, TX 79409

American Psychology Association, Psychology Abstracts, Executive Editor,
1200 17th St NW, Washington, DC 20036

CO, Navy Submarine Base New London, ATTN: Psychology Section, Box
00, Groton, CT 06340

Scientific Technical Information Office, NASA, Washington, DC 20546

Director Defense Research - Engineering, ARPA, Behavioral Science
Division, ATTN: LCOL A. W. Kibler, Washington, DC 20301

HUMAN FACTORS DISTRIBUTION LIST (CONT)

Naval Aerospace Medical Institute, NAVAEROSPREGMEDCEN, ATTN:
Chief Aviation Psychology Division, Pensacola, FL 32512

CO, Naval Health Research Center, San Diego, CA 92152

Commander, Naval Air Systems Command, Code 03, Washington, DC 20361

Commander, Naval Sea Systems Command 047C12, ATTN: CDR George
N. Graine, Washington, DC 20362

Commander, Naval Electronic Systems Command, Code 03, Washington,
DC 20360

Commander, Naval Supply Systems Command, Code 03, Washington,
DC 20376

Commander, Naval Sea Systems Command, Code 03, Washington, DC 20360

Commander, Naval Air Development Center, ATTN: Human Engineering
Branch, Warminster, PA 18974

Human Factors Engineering Division, NAVAIRDEVCECEN, Code 4024, ATTN:
LCDR Charles Thelsen, Warminster, PA 18974

CO, PAC MISS TEST CTR, ATTN: Hd Human Factors, Engineering Branch,
Pt Mugu, CA 93042

Chief Naval Reserve, Code 02, New Orleans, LA 70146

Chief Naval Education and Training, N-3, ATTN: Capt A. McMichael,
NAS, Pensacola, FL 32508

Chief Naval Education and Training, ATTN: B. C. Stone, NAS, Pensacola,
FL 32508

Chief Naval Education and Training, Code 01A, ATTN: Dr W. Maloy, NAS,
Pensacola, FL 32508

CO, Naval Air Technical Training, ATTN: Dr G. D. Mayc, Hd, Research
Branch, NAS Memphis, Millington, TN 38054

Chief Naval Training, ATTN: J. L. Lantoski, NAS, Corpus Christi, TX
78419

HUMAN FACTORS DISTRIBUTION LIST (CONT)

Chief Naval Technical Training, Code 34, ATTN: Dr Harding, NAS
Memphis 75, Millington, TN 38054

CO, NAVED TRAINSUPPCENPAC, Fleet Station PO Bldg, Code N, ATTN:
Mr. Rothenberg, San Diego, CA 92132

Chief Naval Education and Training Support, Code N-2, Bldg 45, ATTN:
Dr Charles Havens, NAS, Pensacola, FL 32509

Chief Naval Education and Training Support, Code N-241, NAS, Pensacola,
FL 32509

US Air Force Human Relations Lab, AFHRL-AS, Advance Systems Division,
Wright-Patterson AFB, OH 45433

US Air Force Human Relations Lab, AFHRL/OR Occupational Manpower
Relations Division, Lackland AFB, TX 78236

US Air Force Human Relations Lab, AFHRL/SM Computational Sciences
Division, Lackland AFB, TX 78235

HQ, Air Training Command, XPT, ATTN: Dr John Meyer, Randolph AFB,
TX 78148

US Air Force Human Relations Lab/DOJZ, Brooks AFB, TX 78235

Chief, Institute Technical Division, ADC DOTI, ATTN: Mr R. E. Coward,
Ent AFB, CO 80912

HQ, US Air Force Systems Command, DLSL, Office Scientific Research,
Andrews AFB, Washington, DC 20331

US Air Force Human Relations Lab, AFHRL-TT, Technical Training
Division, Lowry AFB, CO 80230

US Air Force Human Relations Lab, AFHRL-FT, Flying Training Division,
Williams AFB, AZ 85224

ASD SMSE, ATTN: Mr Harold Kottmann, Wright-Patterson AFB, OH
45433

ENCT, ATTN: Mr Arthur Doty, Wright-Patterson AFB, Dayton, OH 45433

HUMAN FACTORS DISTRIBUTION LIST (CONT)

Commander, Navy Air Force, US Pacific Fleet, NAS North Island, San Diego, CA 92135

Commander, Training Command, ATTN: Educational Advisor, US Pacific Fleet, San Diego, CA 92147

Commander, Training Command, ATTN: Educational Advisor, US Atlantic Fleet, Norfolk, VA 23511

USAF Human Relations Lab, Personnel Research Division, Lackland AFB, TX 78236

DISTRIBUTION LIST

Defense Documentation Center 12
Cameron Station
Alexandria, VA 22314

Naval Training Equipment Center 126
Orlando, FL 32813